Inverse-Variance Weighting for Estimation of Heterogeneous Treatment Effects

Aaron Fisher [1]

Abstract

Many methods for estimating conditional average treatment effects (CATEs) can be expressed as weighted pseudo-outcome regressions (PORs). Previous comparisons of POR techniques have paid careful attention to the choice of pseudo-outcome transformation. However, we argue that the dominant driver of performance is actually the choice of *weights*. For example, we point out that R-Learning implicitly performs a POR with *inverse-variance weights* (IVWs). In the CATE setting, IVWs mitigate the instability associated with inverse-propensity weights, and lead to convenient simplifications of bias terms. We demonstrate the superior performance of IVWs in simulations, and derive convergence rates for IVWs that are, to our knowledge, the fastest yet shown without assuming knowledge of the covariate distribution.

1. Introduction

Estimates of conditional average treatment effects (CATEs) allow for treatment decisions to be tailored to the individual. Formally, let $A \in \{0,1\}$ be a binary treatment, let $X \in \mathcal{X}$ be a vector of confounders and treatment effect modifiers, let $Y_{(a)}$ be the potential outcome under treatment a, and let $Y = AY_{(1)} + (1-A)Y_{(0)}$ be the observed outcome. The CATE is defined as $\tau(X) := \mathbb{E}\left(Y_{(1)} - Y_{(0)}|X\right)$. Under conventional assumptions of exchangeability (i.e., $Y_{(1)}, Y_{(0)} \perp A|X$) and positivity (i.e., $\Pr(A = 1|X) \in (c, 1-c)$ for some $c \in (0,1)$), the CATE can be identified as $\tau(x) = \mathbb{E}(Y|X, A=1) - \mathbb{E}(Y|X, A=0)$.

CATE estimation has a rich history going back several decades (see, e.g., Robins & Rotnitzky, 1995; Hill, 2011; Zhao et al., 2012; Imai & Ratkovic, 2013; Hahn et al., 2020; Athey & Imbens, 2016). We focus here on two general approaches: pseudo-outcome regression (POR) and R-learning. Both approaches easily accommodate flexible machine learning tools, and can attain double robustness (DR) properties similar to those established in the average treatment effect (ATE) literature (Nie & Wager, 2020; Kennedy, 2023; see also Scharfstein et al. 1999; Robins et al. 2000; Bang & Robins 2005; Chernozhukov et al. 2022b; Kennedy 2022)

POR aims to derive a noisy but unbiased approximation of $Y_{(1)} - Y_{(0)}$, and to fit a regression to predict this approximation using X (Rubin & van der Laan, 2005; van der Laan, 2006; Tian et al., 2014; Chen et al., 2017; Künzel et al., 2019; Semenova & Chernozhukov, 2020; Curth & van der Schaar, 2021; Foster & Syrgkanis, 2023; see also Buckley & James 1979; Fan & Gijbels 1994; Rubin & van der Laan 2007; Díaz et al. 2018). The approximation of $Y_{(1)} - Y_{(0)}$ is referred to as a "pseudo-outcome," or "unbiasing transformation," because it serves as an observed stand-in for the latent outcome of interest $Y_{(1)} - Y_{(0)}$. For example, if the propensity scores $\mathbb{E}(A|X)$ are known, then an appropriate pseudo-outcome can be derived using inverse propensity weights: $f_{\text{IPW}}(A,Y) := AY/\mathbb{E}(A|X) - (1-A)Y/\mathbb{E}(1-A|X)$. Since $\mathbb{E}\left(f_{\text{IPW}}(A,Y)|X\right) = \tau(X)$, regressing the pseudo-outcomes $f_{\text{IPW}}(A,Y)$ against X produces a sensible estimate of τ (Powers et al., 2018). This regression can be done with any off-the-shelf machine learning algorithm. For this reason, POR methods are sometimes referred to as "meta-algorithms" (Kennedy, 2023).

R-learning estimates the CATE using a moment condition derived by Robinson (1988; see Section 5.2 of Robins et al., 2008; Nie & Wager, 2020; Zhao et al., 2022; Semenova et al., 2023; Kennedy, 2023; Kennedy et al., 2024). While R-Learning is sometimes described as separate from POR, it can also be expressed as a *weighted* POR (Schuler et al., 2018; Knaus et al., 2021; see Section 1.1 for details).

This parallel between R-learning and weighted POR invites the question of whether or not weights should be used in POR more broadly, and, if so, what choice of weights is optimal? In other words, even after confounding bias has been accounted for through a pseudo-outcome transformation (e.g., f_{IPW}), should *additional* weights be used to prioritize the fit of τ of different subregions of $\mathcal{X}$? We aim to shed light on this question through a combination of simulation & theory.

[1]Genentech, Boston, MA, United States. Correspondence to: Aaron Fisher <afishe27@alumni.jh.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

Contribution Summary

The main intuition of this manuscript is that pseudo-outcomes based on inverse-propensity weights are effective at removing confounding, but can be unstable in the face of propensity scores close to zero or one. Inverse-*variance* weights restabilize the POR without reintroducing confounding, since the CATE estimand is conditional on X, and Y is unconfounded within strata of X. This form of reweighting is done implicitly by the R-Learner.

Section 1.1 discusses the above intuition in more detail. Section 2 shows that the intuition bears out in simulations, for several types of pseudo-outcome transformations. Section 3 demonstrates how the framework of weighted POR can be used to study bias terms for CATE estimates, and to derive fast convergence rates. We close with a discussion.

1.1. Stabilizing Weights in CATE Estimation

In this section we outline POR, R-Learning, and inverse-variance weighting (IVW) in more detail. Let $Z := (Y, X, A)$, and let

$$\begin{aligned}\mu_a(X) &= \mathbb{E}\left(Y|X, A=a\right),\\ \eta(X) &= \mathbb{E}\left(Y|X\right),\\ \pi(X) &= \Pr\left(A=1|X\right),\\ \kappa(X) &= \Pr\left(A=0|X\right), \text{ and}\\ \nu(X) &= Var(A|X).\end{aligned}$$

Let $\theta = \{\mu_1, \mu_0, \eta, \pi, \kappa, \nu\}$ denote the full vector of nuisance functions, and let $\hat{\theta} = \{\hat{\mu}_1, \hat{\mu}_0, \hat{\eta}, \hat{\pi}, \hat{\kappa}, \hat{\nu}\}$ be a set of corresponding nuisance estimates. We use μ and $\hat{\mu}$ as shorthand for $\{\mu_0, \mu_1\}$ and $\{\hat{\mu}_0, \hat{\mu}_1\}$ respectively. One of the reasons we include the redundant representations $\pi(x)$ and $\kappa(x) = 1 - \pi(x)$ is to simplify certain formulas and bias results later on (e.g., Eq (6)). The notation "kappa" is meant to be reminiscent of the term "***c***ontrol."

1.1.1. Weights Used in R-Learning

Given a pair of pre-estimated nuisance functions $\hat{\eta}$ and $\hat{\pi}$, the R-Learning estimate of the CATE (τ) is typically written as

$$\arg\min_{\hat{\tau}} \sum_{i=1}^{n} \left[\{Y_i - \hat{\eta}(X_i)\} - \{A_i - \hat{\pi}(X_i)\}\,\hat{\tau}(X_i)\right]^2. \tag{1}$$

The procedure is motivated by the fact that the term in square brackets has mean zero when $\hat{\eta} = \eta$, $\hat{\pi} = \pi$ and $\hat{\tau} = \tau$ (Robinson, 1988). The nuisance estimates $\hat{\eta}$ and $\hat{\pi}$, are typically obtained via *cross-fitting* (CF): splitting the sample into two partitions, using one to estimate $\hat{\eta}$ and $\hat{\pi}$, and using the other to create the summands in Eq (1) (Nie & Wager, 2020; Kennedy et al., 2020; Kennedy, 2022; Chernozhukov et al., 2022a;b; see also related work from, e.g., Bickel 1982; Schick 1986; Bickel & Ritov 1988, as well as Athey & Imbens 2016). In general, we assume in this section that $\hat{\theta}$ is pre-estimated from an independent dataset or sample partition.

A known but often overlooked fact is that the minimization in Eq (1) can equivalently be solved by fitting a *weighted regression* using X to predict

$$f_{\mathrm{U},\hat{\theta}}(Z) := \frac{Y - \hat{\eta}(X)}{A - \hat{\pi}(X)} \tag{2}$$

with weights $\{A - \hat{\pi}(X)\}^2$ and the squared error loss function (see, e.g., Schuler et al., 2018; Knaus et al., 2021; Zhao et al., 2022; Curth & Van Der Schaar, 2023; and the NonParamDML method in the EconML package, Syrgkanis et al. 2021). For this reason, R-Learning is closely related to "U-Learning," a method that fits an *unweighted* regression to predict $f_{\mathrm{U},\hat{\theta}}(Z)$ from X (see the Appendix of Künzel et al., 2019). The motivation for U-Learning is that, if $\hat{\pi} = \pi$ and $\hat{\eta} = \eta$, then $f_{U,\hat{\theta}}$ is a pseudo-outcome in the sense that $\mathbb{E}\left[f_{\mathrm{U},\hat{\theta}}(Z)|X\right] = \tau(X)$ (Robinson, 1988; Künzel et al., 2019; Nie & Wager, 2020).[1] Thus, U-Learning gives an alternative motivation for R-Learning.

Moreover, we can motivate the R-Learner's weights by appealing to the intuition of inverse-variance weighted least squares. We show in Appendix E that, if $\hat{\theta} = \theta$, the treatment effect is null (i.e., $A \perp Y|X$), and the outcome Y is homoskedastic (i.e., $Var(Y|X) = \sigma^2$ is constant), then the pseudo-outcome $f_{U,\hat{\theta}}$ used in R-Learning has conditional variance

$$Var\left(\frac{Y - \eta(X)}{A - \pi(X)}|X\right) \propto \mathbb{E}\left[\left(A - \pi(X)\right)^{-2}|X\right]. \tag{3}$$

In this way, the $\{A - \hat{\pi}(X)\}^2$ weights used by R-Learning are approximate IVWs, and we would expect them to stabilize the regression.

Indeed, Nie & Wager (2020) remark that U-Learning suffers from instability due to the denominator in $f_{\mathrm{U},\hat{\theta}}(Z)$. They find that R-Learning generally outperforms the U-Learner in simulations. Since the R-Learner is equivalent to a weighted U-Learner, this finding effectively means that the $\{A - \hat{\pi}(X)\}^2$ weights used in R-Learning counteract the instabilities of U-Learning. To our knowledge, the implicit connections between R-Learning, U-Learning and IVW have not been discussed in the literature.

Figure 1 shows a simple simulated illustration of how the R-Learner's weights provide stabilization. Here, $X \sim U(0.05, 0.95)$, $\pi(X) = X$, and $Y \sim N(0, 1)$ regardless of the value of (A, X). This implies that $\tau(x) = 0$ for all

[1] This follows from the "Robinson Decomposition."

x, and that the propensity score is most extreme when x is close to 0 or 1. For simplicity of illustration, we briefly assume perfect knowledge of the nuisance functions, and use this knowledge to define pseudo-outcomes according to Eq (2). (We remove this assumption in our theoretical analysis and main simulation study.) Given these pseudo-outcomes, we apply both U-Learning and R-Learning using spline-based, (weighted) POR. We see in Figure 1 that values of x close to 0 or 1 produce extreme propensity scores, which lead to instability in the pseudo-outcomes. While this hinders the U-Learner's performance, the R-Learner is able to provide a more stable result and a lower rMSE by down-weighting observations with extreme propensity scores.

1.1.2. Alternative Motivation for R-Learner's Weights

As an alternative to Eq (3), a similar motivation for the R-Learner's weights can be derived by noting that $\{A-\hat{\pi}(X)\}^2$ is roughly proportional the inverse variance of $f_{\text{U},\hat{\theta}}(Z)$ conditional on *conditional on $\hat{\theta}$, X and A*. More specifically, if $Var\left(Y|A,X\right)=\sigma^2$ is constant, then

$$Var\left(\frac{Y-\hat{\eta}(X)}{A-\hat{\pi}(X)}|A,X,\hat{\theta}\right)\propto\left\{A-\hat{\pi}(X)\right\}^2.$$

Thus, if we were to expand R-Learning to predict $f_{\text{U},\hat{\theta}}$ as a function of *both X and A*, and if $Var(Y|A,X)$ were constant, then $\{A-\hat{\pi}(X)\}^2$ would form appropriate inverse variance weights, producing the regression problem

$$\arg\min_{\hat{g}}\sum_{i=1}^{n}\left\{A_i-\hat{\pi}(X_i)\right\}^2\left\{\frac{Y-\hat{\eta}(X)}{A-\hat{\pi}(X)}-\hat{g}(A_i,X_i)\right\}^2. \tag{4}$$

The change to include A as a covariate is balanced by the fact that, if $\hat{\theta}=\theta$, then the population minimizer for Eq (4), $\mathbb{E}\left[\frac{Y-\eta(X)}{A-\pi(X)}|A,X\right]$, does not actually depend on A. More specifically, the Robinson Decomposition implies that $\mathbb{E}\left[\frac{Y-\eta(X)}{A-\pi(X)}|A,X\right]=\tau(X)$. Reflecting this fact, if we additionally require the solution to Eq (4) to not depend on A, then we recover R-Learning exactly. Again, R-Learning emerges as a form of (constrained) U-Learning with stabilizing weights.

1.1.3. Weights for Covariance-Based Pseudo-Outcomes

A similar connection to stabilizing weights can be seen in the "oracle" version of R-Learning studied by Kennedy (2023; see their Section 7.6.1). This hypothetical oracle model fits a weighted POR to predict the latent function

$$\begin{aligned}f_{\text{cov},\theta}(Z)&:=\frac{\{A-\pi(X)\}\{Y-\eta(X)\}}{\pi(X)\{1-\pi(X)\}}\\&\approx\frac{\{A-\hat{\pi}(X)\}\{Y-\hat{\eta}(X)\}}{\{A-\hat{\pi}(X)\}^2}\\&=f_{\text{U},\hat{\theta}}(A,X,Y),\end{aligned}$$

with weights $\nu(X)=Var(A|X)$. Above, the approximation simply reflects the fact that if $\hat{\pi}=\pi$ then the conditional expectation of the denominators are identical. We refer to $f_{\text{cov},\theta}$ as the "covariance-based" pseudo-outcome, since the expected value of its numerator is $\mathbb{E}\left[Cov\left(A,Y|X\right)\right]$. Again, if the treatment effect is null ($A\perp Y|X$) and the conditional variance of Y is constant (i.e., $Var(Y|X)=\sigma^2$), then

$$Var(f_{\text{cov},\theta}(A,X,Y)|X)\propto\nu(X)^{-1}$$

(see Appendix E). Thus, in the null setting, the oracle R-Learner is an inverse-variance weighted POR.

1.1.4. Weights for the DR-Learner

Another pseudo-outcome transformation that can suffer from instability is the "DR-Learner" (Kennedy, 2023). This method fits a regression using X to predict $f_{\text{DR},\hat{\theta}}(Z)=f_{1,\hat{\theta}}(Z)-f_{0,\hat{\theta}}(Z)$, where

$$\begin{aligned}&f_{a,\hat{\theta}}(Z)\\&\quad=\hat{\mu}_a(X)+\frac{1(A=a)}{a\hat{\pi}(X)+(1-a)\hat{\kappa}(X)}\left(Y-\hat{\mu}_a(X)\right).\end{aligned} \tag{5}$$

If $Var(Y|X,A)=\sigma^2$ is constant, then it is fairly straightforward to show that $Var\left(f_{\text{DR},\hat{\theta}}(Z)|X,\hat{\theta}=\theta\right)=\kappa(X)^{-1}\pi(X)^{-1}\sigma^2$ (Appendix E). Thus, extreme values of the propensity score again lead to regions where the pseudo-outcome has a high variance.

Inspired by this fact, we will see in the sections below that using weights $\hat{\kappa}(X)\hat{\pi}(X)$ when fitting a POR to predict $f_{\text{DR},\hat{\theta}}(Z)$ leads to fast convergence rates and better simulated errors.

Table 1 summarizes the above relationships.

2. Simulations

The goal of this simulation section is to examine the role of weights in POR. We include 6 simulation scenarios, labeled A, B, C, D, E & F. The first four are experiments taken from Nie & Wager (2020), with $|X|$ set equal to 10. Setting E is the "low dimensional" simulated example from Kennedy (2023). Setting F is the simple illustrative example from

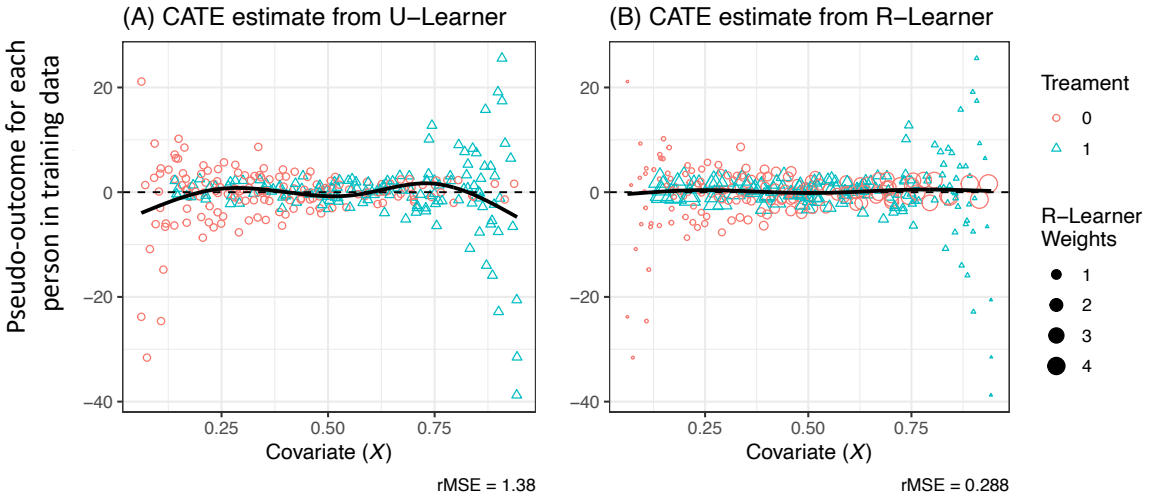

Figure 1. Example of how weights stabilize pseudo-outcome regression, using a single simulated sample. Here, the true conditional average treatment effect is zero for all patients. The estimates from U-Learning & R-Learning are shown as black lines. By down-weighting the observations with high variance, i.e., those with extreme propensity scores, R-Learning is able to achieve a lower rMSE.

Table 1. Different available pseudo-outcome transformations and their conditional variances given X, under certain simplifying assumptions (see Appendix E).

Label	Outcome Transformation	Conditional Variance
U, R ($f_{\mathrm{U},\theta}$)	$\frac{Y-\eta(X)}{A-\pi(X)}$	$\propto \mathbb{E}\left[(A-\pi(X))^{-2}\,\middle|\,X\right]$
DR ($f_{\mathrm{DR},\theta}$)	$\mu_1(X)-\mu_0(X)$ $+\frac{A-\pi(X)}{\pi(X)(1-\pi(X))}\left(Y-\mu_A(X)\right)$	$\propto 1/\nu(X)$
Covariance-based ($f_{\mathrm{cov},\theta}$)	$\frac{\{A-\pi(X)\}\{Y-\eta(X)\}}{\pi(X)(1-\pi(X))}$	$\propto 1/\nu(X)$

Table 2. Simulation Setting Details. Below we show the covariate distribution, CATE function, and nuisance functions for simulations A through F. The notation $\mathrm{trim}_a(b)$ is shorthand for $\min(\max(a,b),1-a)$, and the notation $(a)_+$ is shorthand for $\max(a,0)$. Settings A-D use multivariate, iid covariates X with a dimension of 10. Here, each element of X follows the distribution shown in the second column. Simulations E & F use univariate X. A qualitative description of these simulation settings is shown in Table 3.

Label	X distr.	$\tau(x)$	$\mathbb{E}[Y\|X=x]$	$\mathbb{E}[A\|X=x]$
A	$U(0,1)$	$\frac{1}{2}x_1+\frac{1}{2}x_2$	$\sin(\pi x_1 x_2)+2\left(x_3-\frac{1}{2}\right)^2$	$\mathrm{trim}_{0.1}\{\sin(\pi x_1 x_2)\}$
B	$N(0,1)$	$\log(1+e^{x_2})$ $+x_1$	$\max\{0,x_1+x_2,x_3\}$ $+(x_4+x_5)_+$	$1/2$
C	$N(0,1)$	1	$2\log\left(1+e^{x_1+x_2+x_3}\right)$	$\frac{1}{1+e^{x_2+x_3}}$
D	$N(0,1)$	$\left(\sum_{i=1}^3 x_i\right)_+$ $-(x_4+x_5)_+$	$\left(\sum_{i=1}^3 x_i\right)_+$ $+\frac{1}{2}(x_4+x_5)_+$	$\frac{1}{1+e^{-x_1}+e^{-x_2}}$
E	$U(-1,1)$	0	$1(x_1\le -.5)\frac{(x_1+2)^2}{2}$ $+1(x_1>.5)(x_1+0.125)$ $+\left(\frac{x_1}{2}+0.875\right)1\left(-\frac{1}{2}<x_1<0\right)$ $+\left\{1\left(0<x_1<\frac{1}{2}\right)\right.$ $\left.\times\left(-5\left(x_1-\frac{1}{5}\right)^2+1.075\right)\right\}$	$0.1+(0.8x_1)_+$
F	$U\left(\frac{1}{20},\frac{19}{20}\right)$	0	1	x_1

Table 3. Qualitative summary of the simulation settings detailed in Table 2.

Label	Description	$\tau(x)$	$\mathbb{E}[Y\|X=x]$	$\mathbb{E}[A\|X=x]$
A	Simple effect	Simple	Complex	Complex
B	Randomized trial	Moderate	Moderate	Constant
C	Complex prognosis	Constant	Complex	Simple
D	Unrelated arms	Moderate	Moderate	Moderate
E	Non-differentiable prognosis	Constant	Complex	Simple
F	Simple illustration	Constant	Constant	Simple

Figure 1. Table 2 presents each setting in detail, and Table 3 gives a qualitative summary of each setting. The settings generally differ in their complexity for the functions η, τ and π. In each setting, we simulated sample sizes of 250, 500 and 1000.

We implemented POR with three pseudo-outcome functions: $f_{U,\hat{\theta}}$, $f_{\text{DR},\hat{\theta}}$, and $f_{\text{cov},\hat{\theta}}$. In each case we used 10-fold cross-fitting. For example, for $f_{U,\hat{\theta}}$, we used 90% of the data to estimate the nuisance functions $\hat{\theta}$, evaluated and stored $f_{U,\hat{\theta}}(Z_i)$ for the remaining 10%, and then repeated this process 10 times with different fold assignments to obtain a pseudo-outcome for every individual. We then fit a regression against all of these pseudo-outcomes together. We used boosted trees to perform all of our nuisance regressions, as well as the final regression predicting pseudo-outcomes as a function of X.[2]

For each pseudo-outcome function, we considered a weighted and unweighted version. For $f_{U,\hat{\theta}}$ we compare uniform weights (i.e., the U-Learner) against weights $\{A-\hat{\pi}(X)\}^2$ (i.e., the R-Learner). For $f_{\text{DR},\hat{\theta}}$ and $f_{\text{cov},\hat{\theta}}$, we compare uniform against weights $\hat{\pi}(X)(1-\hat{\pi}(X))$ (see Table 1).

As a baseline comparator, we consider a "T-Learner" approach (Künzel et al., 2019), which entails separately fitting two estimates $\hat{\mu}_1$ and $\hat{\mu}_0$ for μ_1 and μ_0 respectively and then taking $\hat{\mu}_1(x_{\text{new}}) - \hat{\mu}_0(x_{\text{new}})$ as an estimate of $\tau(x_{\text{new}})$. We used the same boosted tree algorithm when fitting the T-Learner.

Figure 2 shows the results of 600 simulation iterations. Weighted POR matched or outperformed unweighted POR in every setting. Performance was similar across the three weighted POR methods we considered. The T-Learner performed comparably to weighted POR in Settings D, E & F, but dramatically underperformed in Settings A, B & C.

[2]Specifically, we used the lightgbm R package written by Shi et al. (2023).

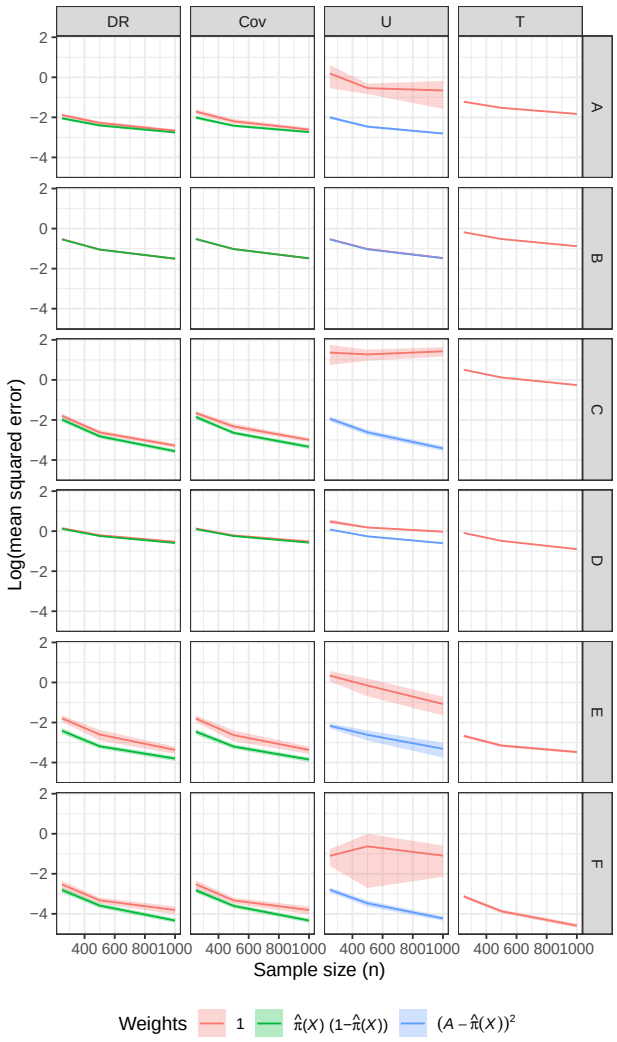

Figure 2. Weighted vs unweighted estimation of simulated CATEs. The first three columns respectively represent POR with the DR-Learner pseudo-outcome ($f_{\text{DR},\hat{\theta}}$), the covariance-based pseudo-outcome ($f_{\text{cov},\hat{\theta}}$), and the U-Learner pseudo-outcome ($f_{\text{U},\hat{\theta}}$). The fourth column shows the T-Learner. The rows show the different simulation settings.

3. Convergence Rate Results

Part of the value the IVW framework is that it provides a straightforward path for simplifying expressions for the bias of CATE estimates. Specifically, if $Z, \hat{\kappa}, \hat{\pi}$, and $\hat{\mu}$ are mutually independent, we can make use of the following helpful identity.

$$\begin{aligned}
&\mathbb{E}\left(\hat{\kappa}\hat{\pi}\left(f_{1,\hat{\theta}} - f_{1,\theta}\right)|X\right) \\
&\quad = \mathbb{E}\left(\hat{\kappa}\hat{\pi}\pi\left(\frac{1}{\hat{\pi}} - \frac{1}{\pi}\right)(\hat{\mu}_1 - \mu_1)|X\right) \\
&\quad = \mathbb{E}\left(\hat{\kappa}|X\right)\mathbb{E}\left(\pi - \hat{\pi}|X\right)\mathbb{E}\left(\hat{\mu}_1 - \mu_1|X\right). \qquad (6)
\end{aligned}$$

The left-hand side is the weighted conditional bias in estimating $f_{1,\hat{\theta}}$ (see Eq (5)), which we can see depends only on the *product* of the biases for $\hat{\pi}$ and $\hat{\mu}$. The first equality is shown in Appendix C. The second comes from the independence assumption. Kennedy (2023) employs a similar identity when reducing bias terms associated with the oracle R-Learner (see their Section 7.6). In the remainder of this section, Eq (6) will play a fundamental role in our study of convergence rates.

3.1. Notation

Let $\bar{\mathbf{Z}} = (\bar{\mathbf{X}}, \bar{\mathbf{a}}, \bar{\mathbf{y}})$ denote a dataset of n observations used for POR, which we assume is independent of the data used for estimating the nuisance functions $\hat{\theta}$. Let d denote the dimension of the domain $\mathcal{X}$ of X, and let x_{new} be a point for which we would like to predict $\tau(x_{\text{new}})$.

We will often use the "bar" notation when referring to estimators derived from $\bar{\mathbf{Z}}$; "hat" notation when referring to quantities that depend on nuisance training data; and both notations when referring to estimators derived from both datasets. We do this to help keep track of dependencies between estimated quantities. Let $\mathbf{X}_{\text{all}}$ be the combined matrix of covariates including $\bar{\mathbf{X}}$ as well as the covariates used in training nuisance functions.

Next we introduce notation to describe convergence rates. From random variables A_n and B_n, let $A_n \lesssim B_n$ denote that there exists a constant c such that $A_n \leq cB_n$ for all n. Let $A_n \asymp B_n$ denote that $A_n \lesssim B_n$ and $B_n \lesssim A_n$. Let $A_n \lesssim_{\mathbb{P}} c_n$ denote that $A_n = O_{\mathbb{P}}(c_n)$ for constants c_n.For any vector k-length vector a, let $\|a\| = \sqrt{\sum_{j=1}^k a_j^2}$ denote its L2 norm.

We say that a function f is s-smooth if there exists a constant c such that $|f(x) - f_{\lfloor s \rfloor, x'}(x)| \leq c||x - x'||^s$ for all x, x', where $\lfloor s \rfloor$ is the largest integer strictly smaller than s and $f_{\lfloor s \rfloor, x'}$ is the $\lfloor s \rfloor^{th}$ order Taylor approximation of f at x'. This form of smoothness is a key property of functions in a Holder class. For completeness, we review this connection in Appendix F.

For any function $g(Z)$, let $\bar{\mathbb{P}}_n(g(Z)) := \frac{1}{n}\sum_{i=1}^n g(Z_i)$ denote its sample average over $\bar{\mathbf{Z}}$. We frequently omit function arguments when clear from context, writing, for example, $\bar{\mathbb{P}}_n(\pi)$ in place of $\bar{\mathbb{P}}_n(\pi(X))$.

3.2. Setup & Assumptions

We study convergence rates for a local polynomial (LP) estimator of $\tau(x_{\text{new}})$ that downweights regions of the covariate space that are either far from x_{new} or that produce pseudo-outcomes with high variance. More formally, let h be a bandwidth parameter that we expect will shrink with n and let $\texttt{kern}$ be a kernel function that is zero outside of the unit hypersphere. Within the unit hypersphere, we assume that $\texttt{kern}$ is bound above and bounded below away from zero. Let $K(X) := \frac{1}{h^d}\texttt{kern}\left(\frac{X - x_{\text{new}}}{h}\right)$. Let $f_{\texttt{basis}}$ be a L-dimensional, polynomial basis function that is bounded on the unit hypersphere, and let $b(X) := f_{\texttt{basis}}\left(\frac{X - x_{\text{new}}}{h}\right)$.

Given independent estimates $\hat{\pi}$, $\hat{\kappa}$ and $\hat{\mu}$, we define $\hat{\nu}(X) := \hat{\pi}(X)\hat{\kappa}(X)$, and define our estimate of $\tau(x_{\text{new}})$ as

$$\hat{\bar{\tau}}(x_{\text{new}}) := \frac{1}{n}\sum_{i=1}^n \hat{\bar{w}}(X_i) f_{\text{DR},\hat{\theta}}(Z_i),$$

where

$$\hat{\bar{w}}(x) := b(x_{\text{new}})^\top \hat{\bar{\mathbf{Q}}}^{-1} b(x) K(x) \hat{\nu}(x)$$

and

$$\hat{\bar{\mathbf{Q}}} := \frac{1}{n}\sum_{i=1}^n b(X_i)\hat{\nu}(X_i)K(X_i)b(X_i)^\top.$$

Thus, $\hat{\bar{\tau}}(x_{\text{new}})$ is a weighted LP regression predicting $f_{\text{DR},\hat{\theta}}(Z)$ from X, with stabilizing weights $\hat{\nu}(X)$. Hereafter, with some abuse of notation, we also use the term "weights" to refer to $\hat{\bar{w}}(X)$.

We study $\hat{\bar{\tau}}(x_{\text{new}})$ by comparing it against an oracle counterpart using the same estimated weights $\hat{\bar{w}}$, but using the true function $f_{\text{DR},\theta}$. That is, we define the oracle estimate

$$\hat{\bar{\tau}}_{\text{oracle}}(x_{\text{new}}) := \frac{1}{n}\sum_{i=1}^n \hat{\bar{w}}(X_i) f_{\text{DR},\theta}(Z_i).$$

Given $\hat{\pi}$ and $\hat{\kappa}$, this oracle estimate is a weighted LP regression predicting $f_{\text{DR},\theta}(Z)$ from X, evaluated at the point $X = x_{\text{new}}$.

Next, we present several assumptions. We reuse the notation "c" to refer to generic constants; the same constant need not satisfy all assumptions.

Assumption 3.1. (Regularity) $\mathbb{E}\left(Y^2|A, X\right)$ is bounded.

Assumption 3.2. (Positivity) There exists a constant $c \in (0, 1)$ such that, for all covariate values x, all $a \in \{0, 1\}$, and

all sample sizes n, we have $c \leq \hat{\kappa}(x), \kappa(x), \hat{\pi}(x), \pi(x) < 1 - c$.

Assumption 3.3. (Nuisance Error) There exists a complexity parameter k (e.g., the number of coefficients in a linear model) and constants c, s_μ and s_π, such that, with probability approaching 1, the sequences $\mathsf{V}_{k,n} := ck/n$, $\mathsf{B}_{\pi,k} := ck^{-s_\pi/d}$ and $\mathsf{B}_{\mu,k} := ck^{-s_\mu/d}$ satisfy

$$\begin{aligned} Var(\hat{\pi}(x)|\mathbf{X}_{\text{all}}) &\leq \mathsf{V}_{k,n}, \\ Var(\hat{\kappa}(x)|\mathbf{X}_{\text{all}}) &\leq \mathsf{V}_{k,n}, \\ Var(\hat{\mu}_a(x)|\mathbf{X}_{\text{all}}) &\leq \mathsf{V}_{k,n}, \\ \mathbb{E}(\hat{\pi}(x) - \pi(x)|\mathbf{X}_{\text{all}}) &\leq \mathsf{B}_{\pi,k}, \\ \mathbb{E}(\hat{\kappa}(x) - \kappa(x)|\mathbf{X}_{\text{all}}) &\leq \mathsf{B}_{\pi,k}, \text{ and} \\ \mathbb{E}(\hat{\mu}_a(x) - \mu_a(x)|\mathbf{X}_{\text{all}}) &\leq \mathsf{B}_{\mu,k} \end{aligned}$$

for all x and a. Above, we assume that k grows with n, and that $k < n$.

The bias conditions of Assumption 3.3 will typically require μ_a and π to be s_μ-smooth and s_π-smooth respectively. The variance conditions typically will require the complexity of the nuisance models (i.e., k) to grow at a limited rate. For example, for spline estimators, they generally require the sample covariance matrices to have stable eigenvalues with high probability. This can be ensured by requiring $k \log(k)/n$ to converge zero (see, e.g., Tropp, 2015; Belloni et al., 2015; Newey & Robins, 2018).

Assumption 3.4. (X Distribution) $\mathcal{X}$ is the unit hypersphere; X is continuous; the density of X is bounded above and bounded below away from zero; and $\|x_{\text{new}}\| < 1 - h$.

Assumption 3.4 implies that any point within h distance from x_{new} is in the interior of $\mathcal{X}$, and that $\Pr\left[\|X - x_{\text{new}}\| \leq h\right] \asymp h^d$.

Assumption 3.5. (Limited bandwidth) $nh^d \to \infty$.

Assumptions 3.4 & 3.5 together imply that the expected number of datapoints in an h-size neighborhood around x_{new} is increasing with n.

Assumption 3.6. (Eigenvalue Stability) There exists a constant $c > 0$ such that $\lambda_{\min}\left(\hat{\tilde{\mathbf{Q}}}\right) > c$ with probability approaching 1.

Assumption 3.6 ensures that the weights $\hat{\tilde{w}}$ are bounded in probability. In Appendix G, we show that Assumption 3.6 follows from Assumptions 3.2, 3.4 & 3.5, if we additionally assume that $\mathbb{E}(|\hat{\nu}(x) - \nu(x)|) \to 0$ for all x, and that $\mathbb{E}\left[f_{\texttt{basis}}(U) f_{\texttt{basis}}(U)^\top\right]$ is positive definite, where U is a random variable that is uniformly distributed on the unit hypersphere. This last condition can be ensured by design (see, e.g., Kennedy et al.'s use of Legendre polynomials, as well as Newey & Robins's Assumption 3, 2018, and Kennedy's Theorem 3, 2023).

Assumption 3.7. (Local Nuisance Estimators) There exists a constant c such that $Cov(\hat{\pi}(x), \hat{\pi}(x')) = 0$, $Cov(\hat{\kappa}(x), \hat{\kappa}(x')) = 0$, and $Cov(\hat{\mu}_a(x), \hat{\mu}_a(x')) = 0$ for all x, x', a satisfying $\|x - x'\| > ck^{-1/d}$.

Assumption 3.7 says that the nuisance models' predictions for sufficiently far away points x, x' depend on entirely different training data. This is true, for example, for LP nuisance regression models with a bandwidth $h \propto k^{-1/d}$ (see, e.g., Tsybakov, 2009).

3.3. Convergence Rate Results

The assumptions in the previous section allow us to characterize the difference between $\hat{\tilde{\tau}}(x_{\text{new}})$ and the oracle estimate.

Theorem 3.8. *(Error with respect to oracle) Under Assumptions 3.1-3.7, we have the following results.*

1. *(4-way CF) If $\hat{\pi}, \hat{\kappa}, \hat{\mu}$, and $\bar{\mathbf{Z}}$ are mutually independent, then*

$$\hat{\tilde{\tau}}(x_{new}) - \hat{\tilde{\tau}}_{oracle}(x_{new}) \lesssim_{\mathbb{P}} \sqrt{\frac{1}{nh^d}} + \mathsf{B}_\mu \mathsf{B}_\pi.$$

2. *(3-way CF) If $\hat{\pi}, \hat{\mu}$ and $\bar{\mathbf{Z}}$ are mutually independent; $\hat{\kappa}(x) = 1 - \hat{\pi}(x)$; and $Var\left[\sup_x \left\{\hat{\pi}(x) - \pi(x)\right\}^2 |\mathbf{X}_{all}\right] \lesssim k_n/n$ with probability approaching 1, then*

$$\hat{\tilde{\tau}}(x_{new}) - \hat{\tilde{\tau}}_{oracle}(x_{new}) \lesssim_{\mathbb{P}} \sqrt{\frac{1}{nh^d}} + \mathsf{B}_\mu \left(\mathsf{B}_\pi + \mathsf{V}_{k,n}\right).$$

3. *(2-way CF) If $\{\hat{\pi}, \hat{\mu}\} \perp \bar{\mathbf{Z}}$ and $\hat{\kappa}(x) = 1 - \hat{\pi}(x)$, then*

$$\begin{aligned} &\hat{\tilde{\tau}}(x_{new}) - \hat{\tilde{\tau}}_{oracle}(x_{new}) \\ &\quad \lesssim_{\mathbb{P}} \sqrt{\frac{1}{nh^d}} + \left(\mathsf{B}_\mu + \sqrt{\mathsf{V}_{k,n}}\right)\left(\mathsf{B}_\pi + \sqrt{\mathsf{V}_{k,n}}\right). \end{aligned}$$

The three bounds given by Theorem 3.8 become less powerful as we relax the independence assumptions. As in Newey & Robins (2018) and Kennedy (2023), the independence conditions can be ensured via higher-order cross-fitting, or "nested" cross-fitting, in which separate folds are used to estimate each nuisance function. Higher order cross-fitting is typically impractical in small or moderate sample sizes, as it requires that a smaller fraction of data points be used to train each nuisance function. That said, the effect of dividing our sample into smaller partitions will be asymptotically dwarfed by the effect of a faster convergence rate.

Point 3 makes the weakest assumptions and produces the least powerful bound. It is similar to the bound in Lemma 2 of Nie & Wager, 2020. That is, Point 3 implies that

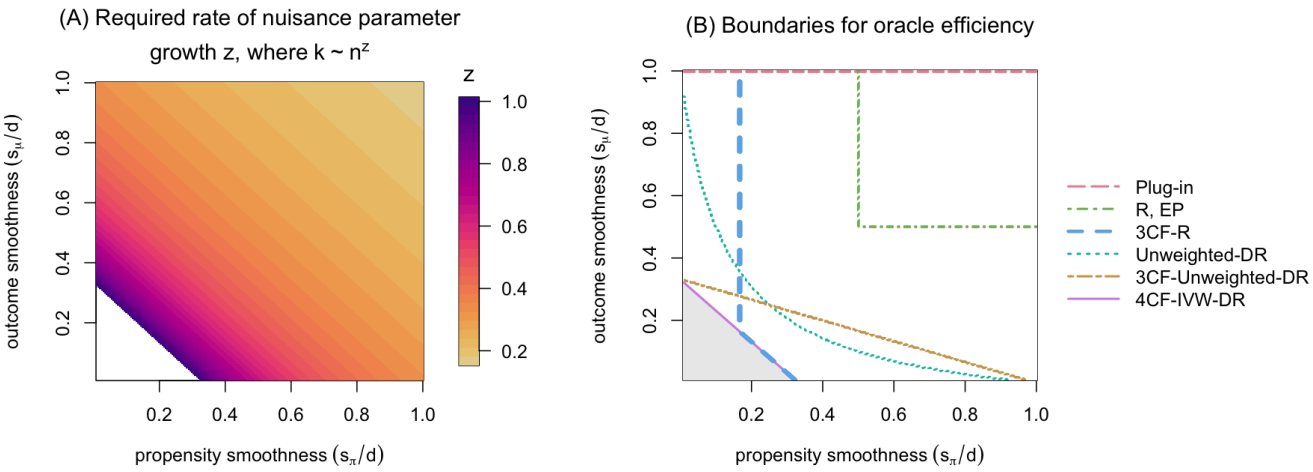

Figure 3. **Panel A** illustrates how increasingly complex nuisance models are required to attain oracle efficiency when the underlying nuisance functions (π, μ) are increasingly non-smooth. For any combination of nuisance smoothness values (s_π, s_μ) and covariate dimension (d), there is a constant z such that the bound from Corollary 3.11 equals the oracle whenever the number of nuisance parameters (k) is proportional to n^z. Panel A shows this rate z using color. For less smooth nuisance functions (the bottom-left of the figure), the value of z becomes higher and higher, until $z \approx 1$ and k grows at a rate almost proportional to the sample size (n). The white triangle in the bottom-left denotes situations where the oracle rate cannot be attained by our approach, or by any other approach. **Panel B** compares this white triangle to related results for other CATE estimators. Each line indicates the boundary of a region within which the estimator has been shown to be oracle efficient. For example, Nie & Wager (2020) show that the standard R-Learner is oracle efficient when s_π/d and s_μ/d fall within the square region on the top-right, marked by the dark green, dot-dashed line. For simplicity, both Panels assume that $s_\tau/d = 1$; Appendix B shows other scenarios.

$\hat{\tilde{\tau}}(x_{\text{new}}) - \hat{\tilde{\tau}}_{\text{oracle}}(x_{\text{new}}) \lesssim_{\mathbb{P}} 1/\sqrt{nh^d}$ if the conditional rMSE of $\hat{\pi}(x)$ and $\hat{\mu}_a(x)$ are $\lesssim n^{-1/4}$. The $\sqrt{1/nh^d}$ term common to all three bounds is a standard variance term associated with LP regression (see, e.g., Proposition 1.13 of Tsybakov, 2009, or Theorem 3 of Kennedy, 2023). The variance condition in Point 2 is similar to Assumption 3.3, and we expect it to hold in similar situations.

To bound the error of the oracle itself, we additionally assume the following.

Assumption 3.9. The target function τ is s_τ-smooth, and the basis $f_{\texttt{basis}}$ is of order at least $\lfloor s_\tau \rfloor$.

From here, fairly standard results for local polynomial regression (e.g., Tsybakov, 2009; see also Kennedy, 2023) imply the following result.

Theorem 3.10. *(Oracle error) Under Assumptions 3.1-3.7 and Assumption 3.9,*

$$\hat{\tilde{\tau}}_{oracle}(x_{new}) - \tau(x_{new}) \lesssim_{\mathbb{P}} \sqrt{\frac{1}{nh^d}} + h^{s_\tau}. \qquad (7)$$

The rate in Theorem 3.10 is minimized when $h \propto n^{\frac{-1}{2s_\tau+d}}$. In this case, the right-hand size of Eq (7) becomes $n^{\frac{-1}{2+d/s_\tau}}$, which is the classic minimax rate for regression with s_τ-smooth functions (see Tsybakov, 2009; and Kennedy et al., 2024). We will say that an estimator is "oracle-efficient" if it achieves this rate.

Combining the results of Theorems 3.8 & 3.10, we see that

Corollary 3.11. *(Final bound) Under Assumptions 3.1-3.7 and Assumption 3.9, if $\hat{\pi}, \hat{\kappa}, \hat{\mu}$ and $\bar{\mathbf{Z}}$ are mutually independent, then*

$$\hat{\tilde{\tau}}(x_{new}) - \tau(x_{new}) \lesssim_{\mathbb{P}} \sqrt{\frac{1}{nh^d}} + h^{s_\tau} + \mathsf{B}_\mu \mathsf{B}_\pi. \qquad (8)$$

Moreover, let $c_{oracle} := \frac{d}{2+d/s_\tau}$. If $s_\pi + s_\mu > c_{oracle}$ and the tuning parameters h and k are selected so that $h \propto n^{\frac{-1}{2s_\tau+d}}$ and $k \propto n^{\frac{d}{(2+d/s_\tau)(s_\pi+s_\mu)}}$, then $\hat{\tilde{\tau}}$ is oracle-efficient:

$$\hat{\tilde{\tau}}(x_{new}) - \tau(x_{new}) \lesssim_{\mathbb{P}} n^{\frac{-1}{2+d/s_\tau}}.$$

The key take-away of Corollary 3.11 is that $\hat{\tilde{\tau}}$ achieves oracle-efficiency when the nuisance functions are sufficiently smooth ($s_\pi + s_\mu > c_{\text{oracle}}$). Importantly, Kennedy et al. (2024) show that it is *impossible* to achieve the oracle rate when $s_\pi + s_\mu < c_{\text{oracle}}$, meaning that $\hat{\tilde{\tau}}$ is oracle-efficient in almost every scenario possible.

The first panel of Figure 3-A illustrates the rate of parameter growth (k) required to reach oracle efficiency in each scenario. This required rate approaches $k \propto n$ as $s_\pi + s_\mu$

approaches c_{oracle}, in the bottom-left. Fittingly, the one exception where we do not attain oracle-efficiency is the edge case where $s_\pi + s_\mu = c_{\text{oracle}}$ exactly, although setting $k \propto n/\log(n)^2$ would still produce the oracle rate up to log factors (as in Kennedy's Corollary 2, 2023). Strictly achieving oracle-efficiency in this edge case would require the number of nuisance parameters in our approach to grow at a rate proportional to n (i.e., $k \propto n$), which will generally preclude Assumption 3.3.

Interestingly, Kennedy (2023) derive a higher-order R-Learner (HORL) that is minimax-optimal even when $s_\pi + s_\mu < c_{\text{oracle}}$ (the white region of Figure 3-A). However, achieving minimax-optimality when $s_\pi + s_\mu < c_{\text{oracle}}$ requires the number of parameters to grow *faster* than the number of relevant data-points. To make this computationally tractable, the HORL assumes external knowledge of the covariate distribution. The HORL can also be used in combination with the empirical distribution of the covariates, but this empirical version leads to non-negligible error terms (see their Proposition 7) and is not necessarily oracle-efficient across the $s_\pi + s_\mu \geq c_{\text{oracle}}$ setting.

To our knowledge, ours is the first estimator shown to be oracle-efficient for any (s_π, s_μ) satisfying $s_\pi + s_\mu < c_{\text{oracle}}$, without assuming knowledge of the covariate distribution. Figure 3-B compares this condition against related smoothness conditions that have previously been shown to imply oracle-efficiency. We compare conditions for: our approach (4CF-IVW-DR); unweighted DR-Learning with 3-way cross-fitting, where the two nuisance models are assumed to be equally complex (3CF-Unweighted-DR; Fisher & Fisher, 2023); unweighted DR-Learning with 2-way cross-fitting, where the two nuisance models can differ in complexity (Unweighted-DR; Kennedy, 2023); R-Learning with 3-way cross-fitting (3CF-R; Kennedy, 2023); R-Learning with 2-way cross-fitting (R; Nie & Wager, 2020); T-Learning, which requires $s_\tau = s_\mu$ ("Plug-in"); and a modified version of T-Learning with stronger guarantees (EP; see Theorem 5 of van der Laan et al., 2024). All of the conditions in Figure 3-B assume s_τ/d is fixed at 1; Appendix B shows how the boundaries change when s_τ varies.

4. Discussion

We have argued that R-Learning implicitly employs a POR with stabilizing weights, and that these weight are key to its success. We also consider doubly robust learners that incorporate IVW more directly, and show that they can attain a convergence rate that is, to our knowledge, the fastest available under our minimal assumptions (Corollary 3.11). An important caveat is that our most powerful results require "higher order sample splitting," which can be impractical in finite samples. With this in mind, we also include weaker results under simpler versions of sample splitting, and study these versions in simulations as well.

The use of weighted regression highlights two fundamental differences in the difficulty of estimating the CATE versus the ATE. The first is that the CATE is inherently a more complex target, and so it incurs a higher oracle error. Indeed, if the underlying CATE function is sufficiently non-smooth, then the oracle error erodes any advantage of using doubly robust methods over plug-in ("T-Learner") methods.

The second is that CATE estimates can incorporate variance-reducing weights without inducing bias, whereas ATE estimates generally cannot. One exception is the special case where the CATE is constant (see, e.g., Hullsiek & Louis, 2002; Yao et al., 2021). More generally though, weighted estimates of an aggregated effect not only change an estimator's precision, but also the estimand itself.

Along these lines, Li et al. (2018) propose a hybrid approach using variance-reducing weights ($\nu(X)$) to estimate the unconditional average effect in the "overlap population," defined as the population who's covariate density d_{overlap} satisfies $d_{\text{overlap}}(x) \propto \nu(x) d_{\text{overall}}(x)$, where d_{overall} is the density of X in the overall population. A drawback of studying the overlap population is that it does not correspond to a specific subgroup of individuals that can be targeted with a deterministic enrollment criteria. On the other hand, Li et al. argue that the overlap population is of special interest due to the fact that it upweights individuals who could plausibly receive either treatment. Morzywolek et al. (2023) extend this idea to the estimation of average treatment effects conditional on a subset of covariates $V \subseteq X$, using weights to emphasize different patients within the strata defined by V. When $V \subset X$, they show that using weights $\nu(X)$ leads to the R Learner objective function. When $V = X$, they note that all weights produce the same estimand (i.e., τ).

The partial aggregation setting where $V \subset X$ is widely relevant, and presents an important area of future research. For example, doctors choosing how to treat patients will not always have access to the same set of confounders X that was used in previous studies. Similarly, policy makers cannot always perfectly tailor intervention programs for each individual. Again, inverse-variance weighted procedures such as R-Learning become harder to apply in this setting without changing the estimand. While this problem can be partially mitigated by fitting an additional regression to predict the R-Learning estimate from a subset of allowed decision factors $V \subset X$, R-Learning may still underperform due to the fact that it internally estimates a target (τ) that is more complex than is necessary (Knaus et al., 2021). Approaches that directly estimate the coarsened function $\mathbb{E}\left(\tau(X)|V\right)$ may improve accuracy due to the low oracle error associated with estimating lower-dimensional functions (see, e.g., Lee et al., 2017; Fisher & Fisher, 2023; as well as Morzywolek et al., 2023).

Acknowledgements

The author is grateful for many conversations with Virginia Fisher that inspired this manuscript, and for her thoughtful comments on early drafts. This work also would not be possible without several helpful conversations with Edward Kennedy. Many of the proofs in this manuscript are based on those shown by Kennedy (2023).

Impact Statement

Conditional effect estimates allow us better understand how any given individual will respond to an intervention. While many existing estimators are unstable in the face of extreme propensity scores, we show that inverse-variance weighting re-stabilizes results and improves performance.

References

Athey, S. and Imbens, G. Recursive partitioning for heterogeneous causal effects. *Proc. Natl. Acad. Sci. U. S. A.*, 113(27):7353–7360, July 2016.

Bang, H. and Robins, J. M. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61 (4):962–973, December 2005.

Belloni, A., Chernozhukov, V., Chetverikov, D., and Kato, K. Some new asymptotic theory for least squares series: Pointwise and uniform results. *J. Econom.*, 186(2):345–366, June 2015.

Bickel, P. J. On adaptive estimation. *Ann. Stat.*, 10(3): 647–671, 1982.

Bickel, P. J. and Ritov, Y. Estimating integrated squared density derivatives: Sharp best order of convergence estimates. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)*, 50(3):381–393, 1988.

Buckley, J. and James, I. Linear regression with censored data. *Biometrika*, 66(3):429–436, 1979.

Chen, S., Tian, L., Cai, T., and Yu, M. A general statistical framework for subgroup identification and comparative treatment scoring. *Biometrics*, 73(4):1199–1209, December 2017.

Chernozhukov, V., Escanciano, J. C., Ichimura, H., Newey, W. K., and Robins, J. M. Locally robust semiparametric estimation. *Econometrica*, 90(4):1501–1535, 2022a.

Chernozhukov, V., Newey, W. K., and Singh, R. Automatic debiased machine learning of causal and structural effects. *Econometrica*, 90(3):967–1027, 2022b.

Curth, A. and van der Schaar, M. Nonparametric estimation of heterogeneous treatment effects: From theory to learning algorithms. In Banerjee, A. and Fukumizu, K. (eds.), *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 1810–1818. PMLR, 2021.

Curth, A. and Van Der Schaar, M. In search of insights, not magic bullets: Towards demystification of the model selection dilemma in heterogeneous treatment effect estimation. In *International Conference on Machine Learning*, pp. 6623–6642. PMLR, 2023.

Díaz, I., Savenkov, O., and Ballman, K. Targeted learning ensembles for optimal individualized treatment rules with time-to-event outcomes. *Biometrika*, 105(3):723–738, September 2018.

Fan, J. and Gijbels, I. Censored regression: Local linear approximations and their applications. *J. Am. Stat. Assoc.*, 89(426):560–570, June 1994.

Fisher, A. and Fisher, V. Three-way Cross-Fitting and Pseudo-Outcome regression for estimation of conditional effects and other linear functionals. June 2023.

Foster, D. J. and Syrgkanis, V. Orthogonal statistical learning. *The Annals of Statistics*, 51(3):879–908, 2023.

Hahn, P. R., Murray, J. S., and Carvalho, C. M. Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion). *Bayesian Analysis*, 15(3):965–1056, 2020.

Hill, J. L. Bayesian nonparametric modeling for causal inference. *J. Comput. Graph. Stat.*, 20(1):217–240, January 2011.

Hullsiek, K. H. and Louis, T. A. Propensity score modeling strategies for the causal analysis of observational data. *Biostatistics*, 3(2):179–193, June 2002.

Imai, K. and Ratkovic, M. Estimating treatment effect heterogeneity in randomized program evaluation. *aoas*, 7 (1):443–470, March 2013.

Kennedy, E. H. Semiparametric doubly robust targeted double machine learning: a review. March 2022.

Kennedy, E. H. Towards optimal doubly robust estimation of heterogeneous causal effects. *EJSS*, 17(2):3008–3049, January 2023.

Kennedy, E. H., Balakrishnan, S., and G'Sell, M. Sharp instruments for classifying compliers and generalizing causal effects. *Ann. Stat.*, 2020.

Kennedy, E. H., Balakrishnan, S., Robins, J. M., and Wasserman, L. Minimax rates for heterogeneous causal effect estimation. *The Annals of Statistics*, 52(2):793–816, 2024.

Knaus, M. C., Lechner, M., and Strittmatter, A. Machine learning estimation of heterogeneous causal effects: Empirical monte carlo evidence. *Econom. J.*, 24(1):134–161, March 2021.

Künzel, S. R., Sekhon, J. S., Bickel, P. J., and Yu, B. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proc. Natl. Acad. Sci. U. S. A.*, 116(10):4156–4165, March 2019.

Lee, S., Okui, R., and Whang, Y.-J. Doubly robust uniform confidence band for the conditional average treatment effect function. *J. Appl. Econ.*, 32(7):1207–1225, November 2017.

Li, F., Morgan, K. L., and Zaslavsky, A. M. Balancing covariates via propensity score weighting. *J. Am. Stat. Assoc.*, 113(521):390–400, January 2018.

Morzywolek, P., Decruyenaere, J., and Vansteelandt, S. On a general class of orthogonal learners for the estimation of heterogeneous treatment effects. March 2023.

Newey, W. K. and Robins, J. R. Cross-Fitting and fast remainder rates for semiparametric estimation. January 2018.

Nie, X. and Wager, S. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 2020.

Powers, S., Qian, J., Jung, K., Schuler, A., Shah, N. H., Hastie, T., and Tibshirani, R. Some methods for heterogeneous treatment effect estimation in high dimensions. *Stat. Med.*, 37(11):1767–1787, May 2018.

Robins, J., Li, L., Tchetgen, E., van der Vaart, A., et al. Higher order influence functions and minimax estimation of nonlinear functionals. In *Probability and statistics: essays in honor of David A. Freedman*, volume 2, pp. 335–422. Institute of Mathematical Statistics, 2008.

Robins, J. M. and Rotnitzky, A. Semiparametric efficiency in multivariate regression models with missing data. *J. Am. Stat. Assoc.*, 90(429):122–129, 1995.

Robins, J. M., Rotnitzky, A., and van der Laan, M. On profile likelihood: Comment. *J. Am. Stat. Assoc.*, 95 (450):477–482, 2000.

Robinson, P. M. Root-N-Consistent semiparametric regression. *Econometrica*, 56(4):931–954, 1988.

Rubin, D. and van der Laan, M. J. A general imputation methodology for nonparametric regression with censored data. 2005.

Rubin, D. and van der Laan, M. J. A doubly robust censoring unbiased transformation. *Int. J. Biostat.*, 3(1), 2007.

Rudelson, M. Random vectors in the isotropic position. *J. Funct. Anal.*, 164(1):60–72, May 1999.

Scharfstein, D. O., Rotnitzky, A., and Robins, J. M. Adjusting for nonignorable Drop-Out using semiparametric nonresponse models. *J. Am. Stat. Assoc.*, 94(448):1096–1120, December 1999.

Schick, A. On asymptotically efficient estimation in semiparametric models. *Ann. Stat.*, 14(3):1139–1151, 1986.

Schuler, A., Baiocchi, M., Tibshirani, R., and Shah, N. A comparison of methods for model selection when estimating individual treatment effects. April 2018.

Semenova, V. and Chernozhukov, V. Debiased machine learning of conditional average treatment effects and other causal functions. *Econom. J.*, 24(2):264–289, August 2020.

Semenova, V., Goldman, M., Chernozhukov, V., and Taddy, M. Inference on heterogeneous treatment effects in high-dimensional dynamic panels under weak dependence. *Quantitative Economics*, 14(2):471–510, 2023.

Serfling, R. J. *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons, 1980.

Shi, Y., Ke, G., Soukhavong, D., Lamb, J., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.-Y., Titov, N., and Cortes, D. lightgbm: Light gradient boosting machine. https://CRAN.R-project.org/package=lightgbm, 2023.

Syrgkanis, V., Lewis, G., Oprescu, M., Hei, M., Battocchi, K., Dillon, E., Pan, J., Wu, Y., Lo, P., Chen, H., Harinen, T., and Lee, J.-Y. Causal inference and machine learning in practice with EconML and CausalML: Industrial use cases at microsoft, TripAdvisor, uber. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, KDD '21, pp. 4072–4073, New York, NY, USA, August 2021. Association for Computing Machinery.

Tian, L., Alizadeh, A. A., Gentles, A. J., and Tibshirani, R. A simple method for estimating interactions between a treatment and a large number of covariates. *J. Am. Stat. Assoc.*, 109(508):1517–1532, October 2014.

Tropp, J. A. An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8 (1-2):1–230, 2015.

Tsybakov, A. B. Introduction to nonparametric estimation. *Springer Series in Statistics*, 2009.

van der Laan, L., Carone, M., and Luedtke, A. Combining t-learning and DR-learning: a framework for oracle-efficient estimation of causal contrasts. February 2024.

van der Laan, M. J. Statistical inference for variable importance. *Int. J. Biostat.*, 2(1), February 2006.

Yao, L., Chu, Z., Li, S., Li, Y., Gao, J., and Zhang, A. A survey on causal inference. *ACM Trans. Knowl. Discov. Data*, 15(5):1–46, May 2021.

Zhao, Q., Small, D. S., and Ertefaie, A. Selective inference for effect modification via the lasso. *J. R. Stat. Soc. Series B Stat. Methodol.*, 84(2):382–413, April 2022.

Zhao, Y., Zeng, D., Rush, A. J., and Kosorok, M. R. Estimating individualized treatment rules using outcome weighted learning. *J. Am. Stat. Assoc.*, 107(449):1106–1118, September 2012.

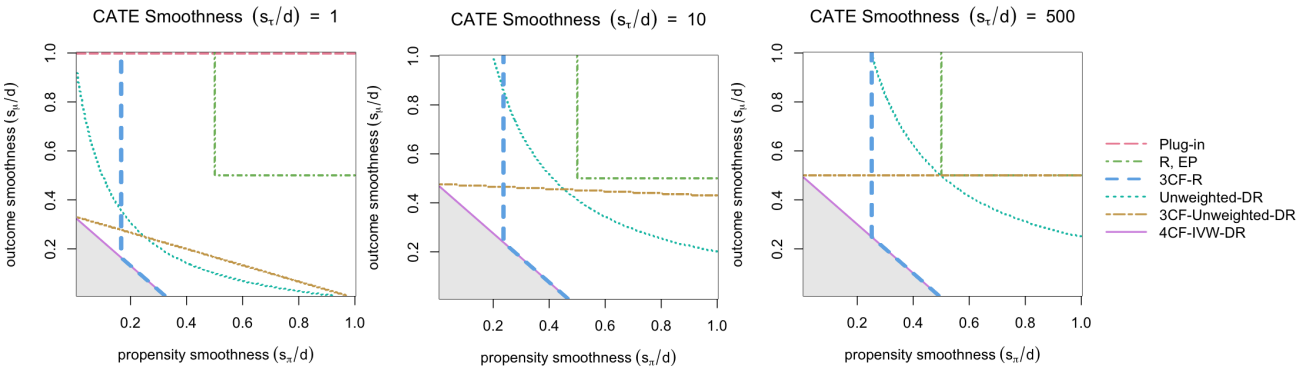

Figure 4. Boundaries for varying CATE smoothness (s_τ/d). Above, we show how the boundaries in Figure 3 change with s_τ/d. Again, each line indicates the boundary of a region within which the estimator has been shown to be oracle efficient. The gray, shaded regions show scenarios where the oracle rate is unattainable.

A. Code

Code for reproducing the methods and simulations in this paper is available at https://github.com/aaronjfisher/wpor/.

B. Theoretical Comparison with Varying s_τ

Figure 4 shows how the oracle-efficiency boundaries from Figure 3 change when the CATE function is more smooth. As s_τ/d approaches infinity, the oracle rate $n^{\frac{-1}{2+d/s_\tau}}$ approaches $n^{-1/2}$. In this setting, oracle efficiency represents a higher performance bar, which can only be attained under stricter conditions on the nuisance smoothness.

C. Proof of Theorem 3.8

Throughout this appendix, we will sometimes use colored text when writing long equations to flag parts of an equation that change from one line to the next (e.g., Line (9)). We use I.E. as an abbreviation for "iterating expectations."

Proof. Throughout the sections below we will use the fact if $1_n A_n \lesssim_{\mathbb{P}} b_n$ and 1_n is an indicator satisfying $\Pr(1_n = 1) \to 1$ (at any rate), then $A_n \lesssim_{\mathbb{P}} b_n$ as well. In particular, we define $\hat{\hat{1}}$ to be the event that the inequalities in Assumptions 3.3 and 3.6 hold. By these same assumptions, $\Pr(\hat{\hat{1}} = 1) \to 1$. When attempting to bound any given term A_n in probability, it will be sufficient to bound $\hat{\hat{1}} A_n$.

We can now present a proof outline. First, we decompose the error with respect to the oracle as

$$\begin{aligned}\hat{\hat{\tau}}(x_{\text{new}}) - \hat{\hat{\tau}}_{\text{oracle}}(x_{\text{new}}) &= \bar{\mathbb{P}}_n \left\{ \hat{\hat{w}} \left(\left(f_{1,\hat{\theta}} - f_{0,\hat{\theta}} \right) - \left(f_{1,\theta} - f_{0,\theta} \right) \right) \right\} \\ &= \bar{\mathbb{P}}_n \left\{ \hat{\hat{w}} \left(f_{1,\hat{\theta}} - f_{1,\theta} \right) \right\} - \bar{\mathbb{P}}_n \left\{ \hat{\hat{w}} \left(f_{0,\hat{\theta}} - f_{0,\theta} \right) \right\}.\end{aligned}$$

Due to the symmetry of the problem, proving that either one of the above terms is bounded will be sufficient. Without loss

of generality, we focus on the first term. After multiplying by $\hat{\mathbb{1}}$, which does not change the bound, we have

$$\begin{aligned}
\hat{\mathbb{1}}\bar{\mathbb{P}}_n\left\{\hat{w}\left(f_{1,\hat{\theta}}-f_{1,\theta}\right)\right\} &= \hat{\mathbb{1}}\bar{\mathbb{P}}_n\left[\hat{w}\left\{\hat{\mu}_1-\mu_1+\frac{A}{\hat{\pi}}\left(Y-\hat{\mu}_1\right)-\frac{A}{\pi}\left(Y-\mu_1\right)\right\}\right] \\
&= \hat{\mathbb{1}}\bar{\mathbb{P}}_n\Bigg[\hat{w}\Bigg\{\hat{\mu}_1-\mu_1-\frac{A}{\pi}\hat{\mu}_1+\frac{A}{\pi}\mu_1 \\
&\quad +\frac{A}{\hat{\pi}}Y-\frac{A}{\hat{\pi}}\mu_1-\frac{A}{\pi}Y+\frac{A}{\pi}\mu_1 \\
&\quad -\frac{A}{\hat{\pi}}\hat{\mu}_1+\frac{A}{\hat{\pi}}\mu_1+\frac{A}{\pi}\hat{\mu}_1-\frac{A}{\pi}\mu_1\Bigg\}\Bigg] \qquad (9)\\
&= \hat{\mathbb{1}}\bar{\mathbb{P}}_n\left[\hat{w}\left(1-\frac{A}{\pi}\right)\left(\hat{\mu}_1-\mu_1\right)\right] \qquad (10)\\
&\quad +\hat{\mathbb{1}}\bar{\mathbb{P}}_n\left[\hat{w}A\left(\frac{1}{\hat{\pi}}-\frac{1}{\pi}\right)\left(Y-\mu_1\right)\right] \qquad (11)\\
&\quad -\hat{\mathbb{1}}\bar{\mathbb{P}}_n\left[\hat{w}A\left(\frac{1}{\hat{\pi}}-\frac{1}{\pi}\right)\left(\hat{\mu}_1-\mu_1\right)\right]. \qquad (12)
\end{aligned}$$

Section C.1, below, shows that the weights $\hat{w}$ satisfy $\mathbb{E}\left(\hat{\mathbb{1}}\hat{w}(X_i)^2\right) \lesssim 1/h^d$ (as in Kennedy (2023)'s Lemma 1). Under the condition that $(\hat{\pi},\hat{\kappa},\hat{\mu}_1) \perp \bar{\mathbf{Z}}$, Section C.2 shows that Lines (10) & (11) are weighted averages of terms that are iid and mean zero, conditional $\hat{\pi},\hat{\kappa},\hat{\mu}_1$ and $\bar{\mathbf{X}}_{\text{all}}$. It will follow that Lines (10) & (11) have expected conditional variance bounded by $1/\left(nh^d\right)$. Thus, Lines (10) & (11) are

$$\lesssim_{\mathbb{P}} \frac{1}{\sqrt{nh^d}} \qquad (13)$$

by Markov's Inequality (see Section C.2 for details). This fact holds for all forms of independence considered in Theorem 3.8 (Points 1, 2 & 3), as it depends only on $(\hat{\pi},\hat{\kappa},\hat{\mu}_1) \perp \bar{\mathbf{Z}}$. As an aside, these same steps can be used to show the first equality in Eq (6).

Line (12) does *not* have mean zero given $\hat{\pi},\hat{\kappa},\hat{\mu}_1$ and $\bar{\mathbf{X}}_{\text{all}}$, and so constitutes the bias relative to the oracle. These terms are more challenging to tackle due to the correlations between the $\hat{\mathbf{Q}}$ matrix (contained within $\hat{w}$) and the $1/\hat{\pi}$ nuisance estimate. However, we can separate these quantities using the Cauchy Schwartz inequality. Line (12) becomes

$$\begin{aligned}
&\hat{\mathbb{1}}\bar{\mathbb{P}}_n\left\{\hat{w}A\left(\frac{1}{\hat{\pi}}-\frac{1}{\pi}\right)\left(\hat{\mu}_1-\mu_1\right)\right\} \\
&= \hat{\mathbb{1}}b(x_{\text{new}})^\top\hat{\mathbf{Q}}^{-1}\bar{\mathbb{P}}_n\left\{b(X_i)K(X_i)\hat{\nu}(X_i)A_i\left(\frac{1}{\hat{\pi}}-\frac{1}{\pi}\right)\left(\hat{\mu}_1-\mu_1\right)\right\} && \text{def of } \hat{w} \\
&\leq \hat{\mathbb{1}}\left\|\hat{\mathbf{Q}}^{-1}b(x_{\text{new}})\right\|\ \left\|\bar{\mathbb{P}}_n\left\{bK\hat{\nu}A\left(\hat{\pi}^{-1}-\pi^{-1}\right)\left(\hat{\mu}_1-\mu_1\right)\right\}\right\| && \text{Cauchy Schwartz} \\
&\lesssim \hat{\mathbb{1}}\left\|\bar{\mathbb{P}}_n\left\{bK\hat{\nu}A\left(\hat{\pi}^{-1}-\pi^{-1}\right)\left(\hat{\mu}_1-\mu_1\right)\right\}\right\| && \text{def of } \hat{\mathbb{1}} \text{ \& } b \\
&= \left[\sum_{l=1}^{L}\hat{\mathbb{1}}\bar{\mathbb{P}}_n\left\{b_\ell K\hat{\nu}A\left(\hat{\pi}^{-1}-\pi^{-1}\right)\left(\hat{\mu}_1-\mu_1\right)\right\}^2\right]^{1/2} \\
&\leq \sum_{l=1}^{L}\left|\hat{\mathbb{1}}\bar{\mathbb{P}}_n\left\{b_\ell K\hat{\kappa}\hat{\pi}A\left(\hat{\pi}^{-1}-\pi^{-1}\right)\left(\hat{\mu}_1-\mu_1\right)\right\}\right|, && (14)
\end{aligned}$$

where the last $\leq$ comes from the definition of $\hat{\nu}$, and from the fact that $\sum_{j=1}^{J}a_j^2 \leq \left(\sum_{j=1}^{J}a_j\right)^2$ for any nonnegative sequences of values $\{a_j,\ldots,a_J\}$.

Appealing to Markov's Inequality, we tackle Line (14) by bounding the second moment of each summand. For Point 1, we

use the fact that $\mathbb{E}(V^2) = Var(V) + \mathbb{E}(V)^2$ for any random variable V to bound

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{E}\left\{\hat{\bar{1}}\bar{\mathbb{P}}_n\left\{b_\ell K\hat{\kappa}\hat{\pi}A\left(\hat{\pi}^{-1}-\pi^{-1}\right)\left(\hat{\mu}_1-\mu_1\right)\right\}^2|\mathbf{X}_{\text{all}},\hat{\kappa}\right\}\right] \\
&= \mathbb{E}\left[\mathbb{E}\left\{\hat{\bar{1}}\bar{\mathbb{P}}_n\left\{b_\ell K\hat{\kappa}\hat{\pi}A\left(\hat{\pi}^{-1}-\pi^{-1}\right)\left(\hat{\mu}_1-\mu_1\right)\right\}|\mathbf{X}_{\text{all}},\hat{\kappa}\right\}^2\right] \quad (15)\\
&\quad + \mathbb{E}\left[Var\left\{\hat{\bar{1}}\bar{\mathbb{P}}_n\left\{b_\ell K\hat{\kappa}\hat{\pi}A\left(\hat{\pi}^{-1}-\pi^{-1}\right)\left(\hat{\mu}_1-\mu_1\right)\right\}|\mathbf{X}_{\text{all}},\hat{\kappa}\right\}\right] \quad (16)
\end{aligned}$$

Section C.3 shows that Line (15) is

$$\lesssim k^{-2(s_\mu+s_\pi)/d}$$

when $\hat{\pi} \perp \hat{\kappa}$, using steps similar to those in Eq (6).

Section C.4 shows that Line (16) is $\lesssim 1/\left(nh^d\right)$. Thus, Eq (14) is

$$\lesssim_{\mathbb{P}} \sqrt{\frac{1}{nh^d}} + k^{-(s_\mu+s_\pi)/d}.$$

This, combined with Line (13), completes the proof of Point 1.

Section C.5 shows that Line (14) is

$$\lesssim_{\mathbb{P}} k^{-(s_\mu+s_\pi)/d} + \frac{k^{1-s_\mu/d}}{n} + \sqrt{\frac{1}{nh^d}}$$

under the conditions of Point 2, and Section C.6 shows that Line (14) is

$$\lesssim_{\mathbb{P}} \frac{k}{n} + \frac{k^{1/2-s_\mu/d}}{\sqrt{n}} + \frac{k^{1/2-s_\pi/d}}{\sqrt{n}} + k^{-(s_\mu+s_\pi)/d}$$

under the conditions of Point 3. This completes the proof for Points 2 & 3. □

C.1. Bound on weights

Here we show results for the weights $\hat{\bar{w}}$. Our approach closely follows classic approaches for LP regression (e.g., Tsybakov, 2009; see also Kennedy, 2023). Let $\mathcal{I}(x) = 1(\|x - x_{\text{new}}\| \leq h)$, so that $K(x) = 0$ and $\hat{\bar{w}}(x) = 0$ whenever $\mathcal{I}(x) = 0$ by the definitions of K and $\hat{\bar{w}}$.

Lemma C.1. *(Bounded weights) Under Assumptions 3.2, 3.4, 3.5, & 3.6:*

1. $K(X) \lesssim \frac{1}{h^d}\mathcal{I}(X)$, *and* $\mathbb{E}\left(K(X)\right) \lesssim \frac{1}{h^d}\mathbb{E}\left(\mathcal{I}(X)\right) \lesssim 1$;
2. $\mathbb{E}\left[\left\{\frac{1}{n}\sum_{i=1}^n K(X_i)\right\}^2\right] \lesssim 1$;
3. $\hat{\bar{1}}|\hat{\bar{w}}(x)| \lesssim \mathcal{I}(x)/h^d$ *for any fixed* x;
4. $\mathbb{E}\left\{\hat{\bar{1}}|\hat{\bar{w}}(X_i)|\right\} \lesssim 1$; *and*
5. $\mathbb{E}\left\{\hat{\bar{1}}\hat{\bar{w}}(X_i)^2\right\} \lesssim 1/h^d$.

Proof. Point 1 comes immediately from the definitions of K and $\mathcal{I}$, and from Assumption 3.4.

For Point 2,

$$\begin{aligned}
\mathbb{E}\left[\left\{\frac{1}{n}\sum_{i=1}^{n}K(X_i)\right\}^2\right] &\lesssim \frac{1}{n^2h^{2d}}\mathbb{E}\left[\left\{\sum_{i=1}^{n}\mathcal{I}(X_i)\right\}^2\right] && \text{Point 1}\\
&= \frac{1}{n^2h^{2d}}\left[\mathbb{E}\left\{\sum_{i=1}^{n}\mathcal{I}(X_i)\right\}+\mathbb{E}\left\{\sum_{i=1}^{n}\mathcal{I}(X_i)\sum_{j\neq i}^{n}\mathbb{E}\left(\mathcal{I}(X_j)|X_i\right)\right\}\right]\\
&\lesssim \frac{1}{n^2h^{2d}}\left[nh^d+n(n-1)h^{2d}\right] && \text{Assm 3.4}\\
&= \frac{1}{nh^d}+\frac{1}{n^2}\left[n(n-1)\right]\\
&\lesssim 1. && \text{Assm 3.5.}
\end{aligned}$$

For Point 3,

$$\begin{aligned}
\hat{\mathbb{1}}|\hat{w}(x)| &\leq \hat{\mathbb{1}}\|b(x_{\text{new}})\|\,\|\hat{\mathbf{Q}}^{-1}b(x)K(x)\hat{\nu}(x)\| && \text{Cauchy Schwartz}\\
&\lesssim \hat{\mathbb{1}}\|\hat{\mathbf{Q}}^{-1}b(x)K(x)\hat{\nu}(x)\| && \text{def of } b\\
&\leq \frac{\hat{\mathbb{1}}}{\lambda_{\min}\left(\hat{\mathbf{Q}}\right)}\|b(x)K(x)\hat{\nu}(x)\|\\
&\lesssim \|b(x)K(x)\hat{\nu}(x)\| && \text{def of } \hat{\mathbb{1}},\ \text{Assm 3.6}\\
&\leq K(x) && \text{def of } b, \text{Assm 3.2}\\
&\lesssim \frac{1}{h^d}\mathcal{I}(x) && \text{Point 1.}
\end{aligned}$$

Point 4 follows from Points 1 & 3. Similarly, for Point 5,

$$\mathbb{E}\left\{\hat{\mathbb{1}}\hat{w}(X_i)^2\right\} \lesssim \frac{1}{h^{2d}}\mathbb{E}\left\{\mathcal{I}(x)\right\} \lesssim \frac{1}{h^d},$$

where the first $\lesssim$ is from Point 3 and the second is from Assumption 3.4. □

C.2. Showing Lines (10) & (11) are $\lesssim_{\mathbb{P}} \sqrt{1/(nh^d)}$

Line (10) has conditional expectation

$$\begin{aligned}
&\hat{\mathbb{1}}\mathbb{E}\left[\bar{\mathbb{P}}_n\left(\hat{\bar{w}}\left(1-\frac{A}{\pi}\right)(\hat{\mu}_1-\mu_1)\right)|\bar{\mathbf{X}}_{\text{all}},\hat{\mu}_1,\hat{\pi},\hat{\kappa}\right]\\
&= \hat{\mathbb{1}}\bar{\mathbb{P}}_n\left(\hat{\bar{w}}\left(1-\frac{\pi}{\pi}\right)(\hat{\mu}_1-\mu_1)\right)\\
&= 0 \qquad (17)
\end{aligned}$$

Thus, the second moment of Line (10) is

$$\begin{aligned}
&\mathbb{E}\left[\hat{\mathbb{1}}\mathbb{E}\left\{\bar{\mathbb{P}}_n\left(\hat{w}\left(1-\frac{A}{\pi}\right)(\hat{\mu}_1-\mu_1)\right)^2|\bar{\mathbf{X}}_{\text{all}},\hat{\mu}_1,\hat{\pi},\hat{\kappa}\right\}\right] \\
&= \mathbb{E}\left[\hat{\mathbb{1}}Var\left\{\bar{\mathbb{P}}_n\left(\hat{w}\left(1-\frac{A}{\pi}\right)(\hat{\mu}_1-\mu_1)\right)|\bar{\mathbf{X}}_{\text{all}},\hat{\mu}_1,\hat{\pi},\hat{\kappa}\right\}\right] && \text{from Eq (17)} \\
&= \mathbb{E}\left[\frac{\hat{\mathbb{1}}}{n^2}\sum_{i=1}^n \hat{w}(X_i)^2(\hat{\mu}_1(X_i)-\mu_1(X_i))^2\frac{1}{\pi(X_i)^2}Var\left[A|\bar{\mathbf{X}}_{\text{all}}\right]\right] \\
&\lesssim \mathbb{E}\left[\frac{\hat{\mathbb{1}}}{n^2}\sum_{i=1}^n \hat{w}(X_i)^2(\hat{\mu}_1(X_i)-\mu_1(X_i))^2\right] && \text{Assm 3.2} \\
&\lesssim \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}\left[\hat{\mathbb{1}}\hat{w}(X_i)^2\mathbb{E}\left\{(\hat{\mu}_1(X_i)-\mu_1(X_i))^2|\mathbf{X}_{\text{all}}\right\}\right] \\
&\lesssim \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}\left[\hat{\mathbb{1}}\hat{w}(X_i)^2\right] && \text{def of } \hat{\mathbb{1}} \\
&\lesssim \frac{1}{nh^d} && \text{Lemma C.1.5.}
\end{aligned}$$

From here, Markov's Inequality implies that Line (10) is $\lesssim_{\mathbb{P}} \sqrt{\frac{1}{nh^d}}$.

Similarly, Line (11) has conditional expectation

$$\begin{aligned}
&\mathbb{E}\left[\hat{\mathbb{1}}\bar{\mathbb{P}}_n\left(\hat{w}A\left(\frac{1}{\hat{\pi}}-\frac{1}{\pi}\right)(Y-\mu_1)\right)|\bar{\mathbf{X}}_{\text{all}},\hat{\mu}_1,\hat{\pi},\hat{\kappa}\right] \\
&= \hat{\mathbb{1}}\bar{\mathbb{P}}_n\left[\hat{w}\left(\frac{1}{\hat{\pi}}-\frac{1}{\pi}\right)\mathbb{E}\left\{A(Y-\mu_1)|X\right\}\right] \\
&= \hat{\mathbb{1}}\bar{\mathbb{P}}_n\left[\hat{w}\left(\frac{1}{\hat{\pi}}-\frac{1}{\pi}\right)\mathbb{E}\left\{Y-\mu_1|X,A=1\right\}\pi(X)\right] \\
&= 0. && (18)
\end{aligned}$$

Thus, the second moment of Line (11) is

$$\begin{aligned}
&\mathbb{E}\left[\hat{\mathbb{1}}\mathbb{E}\left\{\bar{\mathbb{P}}_n\left(\hat{w}A\left(\frac{1}{\hat{\pi}}-\frac{1}{\pi}\right)(Y-\mu_1)\right)^2|\bar{\mathbf{X}}_{\text{all}},\hat{\mu}_1,\hat{\pi},\hat{\kappa}\right\}\right] \\
&= \mathbb{E}\left[\hat{\mathbb{1}}Var\left\{\bar{\mathbb{P}}_n\left(\hat{w}A\left(\frac{1}{\hat{\pi}}-\frac{1}{\pi}\right)(Y-\mu_1)\right)|\bar{\mathbf{X}}_{\text{all}},\hat{\mu}_1,\hat{\pi},\hat{\kappa}\right\}\right] \\
&= \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}\left[\hat{\mathbb{1}}\hat{w}(X_i)^2\left(\frac{1}{\hat{\pi}(X_i)}-\frac{1}{\pi(X_i)}\right)^2 Var\left\{A(Y-\mu_1)|\bar{\mathbf{X}}_{\text{all}}\right\}\right] \\
&\lesssim \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}\left[\hat{\mathbb{1}}\hat{w}(X_i)^2\right] && \text{Assms 3.1 \& 3.2} \\
&\lesssim \frac{1}{nh^d} && \text{Lemma C.1.5.}
\end{aligned}$$

Markov's Inequality then implies that Line (11) is $\lesssim_{\mathbb{P}} \sqrt{\frac{1}{nh^d}}$.

C.3. Showing Line (15) is $\lesssim k^{-2(s_\mu+s_\pi)/d}$ when $\hat{\pi} \perp \hat{\kappa}$

Let $\hat{1}$ be the indicator that the inequalities in Assumption 3.3 hold, where $\hat{1} \geq \hat{\bar{1}}$, and $\hat{1}$ depends only on $\mathbf{X}_{\text{all}}$. The inner expectation in Line (15) equals

$$
\begin{aligned}
&\mathbb{E}\left[\hat{\bar{1}}\bar{\mathbb{P}}_n\left\{b_\ell K\hat{\kappa}\hat{\pi}A\left(\hat{\pi}^{-1}-\pi^{-1}\right)\left(\hat{\mu}_1-\mu_1\right)\right\}|\mathbf{X}_{\text{all}},\hat{\kappa}\right] \\
&\leq \frac{\hat{1}}{n}\sum_{i=1}^n b_\ell(X_i)K(X_i)\hat{\kappa}(X_i) \\
&\qquad \times \mathbb{E}\left[A\left(1-\frac{\hat{\pi}(X_i)}{\pi(X_i)}\right)|\mathbf{X}_{\text{all}}\right]\mathbb{E}\left[\hat{\mu}_1(X_i)-\mu_1(X_i)|\mathbf{X}_{\text{all}}\right] && \text{4-way independence} \qquad (19)\\
&\lesssim \frac{\hat{1}k^{-s_\mu/d}}{n}\sum_{i=1}^n K(X_i)\left|\mathbb{E}\left[A\left(1-\frac{\hat{\pi}(X_i)}{\pi(X_i)}\right)|\mathbf{X}_{\text{all}}\right]\right| && \text{def of } \hat{1} \\
&= \frac{\hat{1}k^{-s_\mu/d}}{n}\sum_{i=1}^n K(X_i)\left|\mathbb{E}\left[\mathbb{E}\left(A|\mathbf{X}_{\text{all}},\hat{\pi}\right)\left(1-\frac{\hat{\pi}(X_i)}{\pi(X_i)}\right)|\mathbf{X}_{\text{all}}\right]\right| && \text{I.E.} \\
&= \frac{\hat{1}k^{-s_\mu/d}}{n}\sum_{i=1}^n K(X_i)\left|\mathbb{E}\left[\pi(X_i)-\hat{\pi}(X_i)|\mathbf{X}_{\text{all}}\right]\right| && \text{by } \mathbb{E}\left(A_i|\mathbf{X}_{\text{all}},\hat{\pi}\right)=\pi(X_i) \\
&\lesssim \frac{k^{-(s_\mu+s_\pi)/d}}{n}\sum_{i=1}^n K(X_i) && \text{def of } \hat{1}.
\end{aligned}
$$

Note that Line (19) requires $\hat{\pi}(x) \perp \hat{\kappa}(x)$ in order to remove the conditioning on $\hat{\kappa}$ from the expectation term containing $\hat{\pi}$.

Thus, Line (15) is

$$
\lesssim k^{-2(s_\mu+s_\pi)/d}\mathbb{E}\left[\left\{\frac{1}{n}\sum_{i=1}^n K(X_i)\right\}^2\right] \lesssim k^{-2(s_\mu+s_\pi)/d}
$$

where the second $\lesssim$ comes from Lemma C.1.2.

C.4. Showing Line (16) is $\lesssim 1/(nh^d)$

Line (16) is the expected value of

$$
\begin{aligned}
&Var\left[\hat{\bar{1}}\bar{\mathbb{P}}_n\left\{b_\ell K\hat{\kappa}A\left(1-\frac{\hat{\pi}}{\pi}\right)\left(\hat{\mu}_1-\mu_1\right)\right\}\;|\;\mathbf{X}_{\text{all}},\hat{\kappa}\right] \\
&= Var\left[\;\mathbb{E}\left[\hat{\bar{1}}\bar{\mathbb{P}}_n\left\{b_\ell K\hat{\kappa}A\left(1-\frac{\hat{\pi}}{\pi}\right)\left(\hat{\mu}_1-\mu_1\right)\right\}\;|\;\mathbf{X}_{\text{all}},\hat{\pi},\hat{\kappa},\hat{\mu}_1\right]\;|\;\mathbf{X}_{\text{all}},\hat{\kappa}\right] \\
&\quad + \mathbb{E}\left[Var\left[\hat{\bar{1}}\bar{\mathbb{P}}_n\left\{b_\ell K\hat{\kappa}A\left(1-\frac{\hat{\pi}}{\pi}\right)\left(\hat{\mu}_1-\mu_1\right)\right\}\;|\;\mathbf{X}_{\text{all}},\hat{\pi},\hat{\kappa},\hat{\mu}_1\right]\;|\;\mathbf{X}_{\text{all}},\hat{\kappa}\right] && \text{Law of Total Var} \\
&= Var\left[\frac{\hat{\bar{1}}}{n}\sum_{i=1}^n b_\ell K\hat{\kappa}(\pi-\hat{\pi})(\hat{\mu}_1-\mu_1)|\mathbf{X}_{\text{all}},\hat{\kappa}\right] && (20)\\
&\quad + \mathbb{E}\left[\frac{\hat{\bar{1}}}{n^2}\sum_{i=1}^n b_\ell^2K^2\hat{\kappa}^2Var(A|\bar{\mathbf{X}})\left(1-\frac{\hat{\pi}}{\pi}\right)^2(\hat{\mu}_1-\mu_1)^2|\mathbf{X}_{\text{all}},\hat{\kappa}\right]. && (21)
\end{aligned}
$$

Section C.4.1 shows that the expectation of Line (20) is $\lesssim 1/(nh^d)$ and Section C.4.2 shows that the expectation of Line (21) is $\lesssim 1/(nh^d)$.

C.4.1. Showing the expectation of Line (20) is $\lesssim 1/(nh^d)$

To study Line (20), it will be helpful to introduce some abbreviations. Let $\epsilon_{\hat{\pi}i} := \hat{\pi}(X_i) - \pi(X_i)$, and $\epsilon_{\hat{\mu}i} := \hat{\mu}_1(X_i) - \mu_1(X_i)$. Line (20) becomes

$$
\begin{aligned}
&Var\left[\frac{\hat{1}}{n}\sum_{i=1}^{n} b_\ell(X_i)K(X_i)\hat{\kappa}(X_i)\epsilon_{\hat{\pi}i}\epsilon_{\hat{\mu}i}|\mathbf{X}_{\text{all}}, \hat{\kappa}_i\right] \\
&\lesssim \frac{\hat{1}}{n^2}\sum_{i=1}^{n} K(X_i)^2 Var\left(\epsilon_{\hat{\pi}i}\epsilon_{\hat{\mu}i}|\mathbf{X}_{\text{all}}\right) &(22)\\
&\quad + \frac{\hat{1}}{n^2}\sum_{i=1}^{n}\sum_{j\in\{1,\dots n\}\setminus i} K(X_i)K(X_j)\ \left|Cov\left(\epsilon_{\hat{\pi}i}\epsilon_{\hat{\mu}i}, \epsilon_{\hat{\pi}j}\epsilon_{\hat{\mu}j}|\mathbf{X}_{\text{all}}\right)\right|, &(23)
\end{aligned}
$$

by the definition of b.

To study these variance and covariance terms, we use the fact that for any four variables A_1, A_2, B_1, B_2 satisfying $(A_1, A_2) \perp (B_1, B_2)$, we have

$$
\begin{aligned}
&Cov(A_1B_1, A_2B_2) \\
&= Cov(A_1, A_2)Cov(B_1, B_2) + \mathbb{E}(A_1)\mathbb{E}(A_2)Cov(B_1, B_2) + Cov(A_1, A_2)\mathbb{E}(B_1)\mathbb{E}(B_2). &(24)
\end{aligned}
$$

A corollary of Eq (24) is that

$$
Var(A_1B_1) = Var(A_1)Var(B_1) + \mathbb{E}(A_1)^2Var(B_1) + Var(A_1)\mathbb{E}(B_1)^2. \tag{25}
$$

Applying Eq (25), we see that Line (22) equals

$$
\begin{aligned}
&\frac{\hat{1}}{n^2}\sum_{i=1}^{n} K(X_i)^2 \{Var(\epsilon_{\hat{\pi}i}|\mathbf{X}_{\text{all}})Var(\epsilon_{\hat{\mu}i}|\mathbf{X}_{\text{all}}) \\
&\qquad +\mathbb{E}(\epsilon_{\hat{\pi}i}|\mathbf{X}_{\text{all}})^2Var(\epsilon_{\hat{\mu}i}|\mathbf{X}_{\text{all}}) + Var(\epsilon_{\hat{\pi}i}|\mathbf{X}_{\text{all}})\mathbb{E}(\epsilon_{\hat{\mu}i}|\mathbf{X}_{\text{all}})^2\} \\
&\lesssim \frac{1}{n^2}\sum_{i=1}^{n} K(X_i)^2 \qquad \text{def of } \hat{1}. &(26)
\end{aligned}
$$

For the off-diagonal terms in Line (23), we first note that for any $i, j \in \{1, \dots n\}$ satisfying $i \neq j$ we have

$$
\begin{aligned}
&\hat{1}Cov\left(\epsilon_{\hat{\pi}i}\epsilon_{\hat{\mu}i}, \epsilon_{\hat{\pi}j}\epsilon_{\hat{\mu}j}|\mathbf{X}_{\text{all}}\right) \\
&= \hat{1}Cov\left(\epsilon_{\hat{\pi}i}, \epsilon_{\hat{\pi}j}|\mathbf{X}_{\text{all}}\right) Cov\left(\epsilon_{\hat{\mu}i}, \epsilon_{\hat{\mu}j}|\mathbf{X}_{\text{all}}\right) \\
&\quad + \hat{1}Cov\left(\epsilon_{\hat{\pi}i}, \epsilon_{\hat{\pi}j}|\mathbf{X}_{\text{all}}\right) \mathbb{E}\left(\epsilon_{\hat{\mu}i}|\mathbf{X}_{\text{all}}\right)^2 + \hat{1}\mathbb{E}\left(\epsilon_{\hat{\pi}i}|\mathbf{X}_{\text{all}}\right)^2 Cov\left(\epsilon_{\hat{\mu}i}, \epsilon_{\hat{\mu}j}|\mathbf{X}_{\text{all}}\right) \qquad \text{by Eq (24)}, \\
&\lesssim \hat{1}Cov\left(\epsilon_{\hat{\pi}i}, \epsilon_{\hat{\pi}j}|\mathbf{X}_{\text{all}}\right) Cov\left(\epsilon_{\hat{\mu}i}, \epsilon_{\hat{\mu}j}|\mathbf{X}_{\text{all}}\right) \\
&\quad + \hat{1}Cov\left(\epsilon_{\hat{\pi}i}, \epsilon_{\hat{\pi}j}|\mathbf{X}_{\text{all}}\right) + \hat{1}Cov\left(\epsilon_{\hat{\mu}i}, \epsilon_{\hat{\mu}j}|\mathbf{X}_{\text{all}}\right) \qquad \text{def of } \hat{1}, &(27)
\end{aligned}
$$

where

$$
\begin{aligned}
&\hat{1}Cov(\epsilon_{\hat{\pi}i}, \epsilon_{\hat{\pi}j}|\mathbf{X}_{\text{all}}) \\
&\quad = \hat{1}Cov(\epsilon_{\hat{\pi}i}, \epsilon_{\hat{\pi}j}|\mathbf{X}_{\text{all}})1\left(\|X_i - X_j\| \leq ck^{-1/d}\right) \qquad \text{Assm 3.7} \\
&\quad \leq \hat{1}Var(\epsilon_{\hat{\pi}i}|\mathbf{X}_{\text{all}})^{1/2}Var(\epsilon_{\hat{\pi}j}|\mathbf{X}_{\text{all}})^{1/2}1\left(\|X_i - X_j\| \leq ck^{-1/d}\right) \qquad \text{Cauchy Schwartz} \\
&\quad \lesssim \frac{k}{n}1\left(\|X_i - X_j\| \leq ck^{-d}\right), \qquad \text{def of } \hat{1}. &(28)
\end{aligned}
$$

By the same reasoning,

$$
\hat{1}Cov(\epsilon_{\hat{\mu}i}, \epsilon_{\hat{\mu}j}|\mathbf{X}_{\text{all}}) \lesssim \frac{k}{n}1\left(\|X_i - X_j\| \leq ck^{-1/d}\right). \tag{29}
$$

Plugging Eqs (28) & (29) into Eq (27), we get

$$\hat{1}Cov\left(\epsilon_{\hat{\pi}i}\epsilon_{\hat{\mu}i}, \epsilon_{\hat{\pi}j}\epsilon_{\hat{\mu}j}|\mathbf{X}_{\text{all}}\right) \lesssim \left(\frac{k^2}{n^2} + 2\frac{k}{n}\right) 1\left(\|X_i - X_j\| \leq ck^{-1/d}\right). \tag{30}$$

Finally, plugging Eqs (26) & (30) into Lines (22) & (23), we see that the expectation of the expectation of Line (22) plus Line (23) is

$$\begin{aligned}
&\lesssim \mathbb{E}\left[\frac{1}{n^2}\sum_{i=1}^{n} K(X_i)^2 \right. \\
&\quad \left. + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\in\{1,\dots n\}\setminus i} K(X_i)K(X_j)\,\frac{k}{n} 1\left(\|X_i - X_j\| \leq ck^{-1/d}\right)\right] \\
&\lesssim \frac{1}{n^2h^{2d}}\sum_{i=1}^{n}\mathbb{E}\left[\mathcal{I}(X_i)\right] \\
&\quad + \frac{k}{n^3h^{2d}}\sum_{i=1}^{n}\sum_{j\in\{1,\dots n\}\setminus i}\mathbb{E}\left[\mathcal{I}(X_i)\mathbb{E}\left\{1\left(\|X_i - X_j\| \leq ck^{-1/d}\right)|X_i\right\}\right] && \text{Lemma C.1.1} \\
&\lesssim \frac{1}{n^2h^{2d}}\sum_{i=1}^{n}\mathbb{E}\left[\mathcal{I}(X_i)\right] \\
&\quad + \frac{k}{n^3h^{2d}}\sum_{i=1}^{n}\sum_{j\in\{1,\dots n\}\setminus i}\mathbb{E}\left[\mathcal{I}(X_i)k^{-1}\right] && \text{Assm 3.4} \\
&\lesssim \frac{1}{nh^d} + \frac{1}{nh^d}. && \text{Lemma C.1.1.}
\end{aligned}$$

Thus, the expectation of Line (20) is $\lesssim 1/(nh^d)$ as well.

C.4.2. Showing the expectation of Line (21) is $\lesssim 1/(nh^d)$

The expectation of Line (21) is

$$\begin{aligned}
&\lesssim \mathbb{E}\mathbb{E}\left[\frac{\hat{1}}{n^2}\sum_{i=1}^{n} K^2\hat{\kappa}^2\left(1 - \frac{\hat{\pi}}{\pi}\right)^2(\hat{\mu}_1 - \mu_1)^2|\mathbf{X}_{\text{all}}, \hat{\kappa}\right] && \text{def of } b \\
&= \mathbb{E}\left[\frac{\hat{1}}{n^2}\sum_{i=1}^{n} K^2\hat{\kappa}^2\mathbb{E}\left\{\left\{\frac{\pi}{\pi}\left(1 - \frac{\hat{\pi}}{\pi}\right)\right\}^2|\mathbf{X}_{\text{all}}\right\}\mathbb{E}\left\{(\hat{\mu}_1 - \mu_1)^2|\mathbf{X}_{\text{all}}\right\}\right] && \text{4-way independence} \\
&= \mathbb{E}\left[\frac{\hat{1}}{n^2}\sum_{i=1}^{n} K^2\hat{\kappa}^2\mathbb{E}\left\{\frac{1}{\pi^2}(\pi - \hat{\pi})^2|\mathbf{X}_{\text{all}}\right\}\mathbb{E}\left\{(\hat{\mu}_1 - \mu_1)^2|\mathbf{X}_{\text{all}}\right\}\right] \\
&\lesssim \mathbb{E}\left[\frac{\hat{1}}{n^2}\sum_{i=1}^{n} K^2\mathbb{E}\left\{(\pi - \hat{\pi})^2|\mathbf{X}_{\text{all}}\right\}\mathbb{E}\left\{(\hat{\mu}_1 - \mu_1)^2|\mathbf{X}_{\text{all}}\right\}\right] && \text{Assm 3.2} \\
&\lesssim \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}\left[K(X_i)^2\right] && \text{def of } \hat{1} \\
&\lesssim \frac{1}{n^2h^{2d}}\sum_{i=1}^{n}\mathbb{E}\left[\mathcal{I}(X_i)\right] && \text{Lemma C.1.1} \\
&\lesssim \frac{1}{nh^d} && \text{Lemma C.1.1.}
\end{aligned}$$

C.5. Bounding Line (14) under the conditions of Point 2

Here, we redefine $\hat{\bar{1}}$ and $\hat{1}$ to additionally indicate that $Var\left[\sup_x \left\{\hat{\pi}(x) - \pi(x)\right\}^2 |\mathbf{X}_{\text{all}}\right] \leq ck_n/n$ for all x. By assumption, we still have $\Pr(\hat{\bar{1}} = 1) \to 1$ and $\Pr(\hat{1} = 1) \to 1$.

We can add and subtract $\kappa(X)$ to see that the summands in Line (14) are

$$\leq \hat{1}|\bar{\mathbb{P}}_n \left\{b_\ell K \left\{\hat{\kappa} - \kappa\right\} \hat{\pi} A \left(\hat{\pi}^{-1} - \pi^{-1}\right) \left(\hat{\mu}_1 - \mu_1\right)\right\}| \tag{31}$$

$$+ \hat{1}|\bar{\mathbb{P}}_n \left\{b_\ell K \kappa \hat{\pi} A \left(\hat{\pi}^{-1} - \pi^{-1}\right) \left(\hat{\mu}_1 - \mu_1\right)\right\}|. \tag{32}$$

Line (32) can be studied in the same way as in Sections C.3 & C.4, producing the same bound. We tackle Line (31) by bounding its second moment, which is equal to

$$\mathbb{E}\left[\quad \mathbb{E}\left\{\hat{1}\bar{\mathbb{P}}_n \left\{b_\ell K \left\{\hat{\kappa} - \kappa\right\} A \left(1 - \frac{\hat{\pi}}{\pi}\right) \left(\hat{\mu}_1 - \mu_1\right)\right\}^2 |\mathbf{X}_{\text{all}}\right\} \right]$$

$$= \mathbb{E}\left[\quad \mathbb{E}\left\{\hat{1}\bar{\mathbb{P}}_n \left\{b_\ell K \left\{\hat{\kappa} - \kappa\right\} A \left(1 - \frac{\hat{\pi}}{\pi}\right) \left(\hat{\mu}_1 - \mu_1\right)\right\} |\mathbf{X}_{\text{all}}\right\}^2 \right] \tag{33}$$

$$+ \mathbb{E}\left[Var\left\{\hat{1}\bar{\mathbb{P}}_n \left\{b_\ell K \left\{\hat{\kappa} - \kappa\right\} A \left(1 - \frac{\hat{\pi}}{\pi}\right) \left(\hat{\mu}_1 - \mu_1\right)\right\} |\mathbf{X}_{\text{all}}\right\} \right]. \tag{34}$$

For Line (33), since $\hat{\kappa}(x) = 1 - \hat{\pi}(x)$, we have

$$\hat{\kappa}(x) - \kappa(x) = 1 - \hat{\pi}(x) - (1 - \pi(x)) = \pi(x) - \hat{\pi}(x),$$

which implies that the inner expectation in Line (33) equals

$$\begin{aligned}
&\frac{\hat{1}}{n}\sum_{i=1}^n b_\ell(X_i) K(X_i) \mathbb{E}\left\{\hat{\mu}_1(X_i) - \mu_1(X_i)|X_i\right\} \\
&\qquad \times \mathbb{E}\left\{\left\{\pi(X_i) - \hat{\pi}(X_i)\right\} A_i \left(1 - \frac{\hat{\pi}(X_i)}{\pi(X_i)}\right) |\mathbf{X}_{\text{all}}\right\} && \hat{\mu} \perp \hat{\pi} \\
&= \frac{\hat{1}}{n}\sum_{i=1}^n b_\ell(X_i) K(X_i) \mathbb{E}\left\{\hat{\mu}_1(X_i) - \mu_1(X_i)|X_i\right\} \\
&\qquad \times \mathbb{E}\left\{\left\{\pi(X_i) - \hat{\pi}(X_i)\right\}^2 |\mathbf{X}_{\text{all}}\right\} && \text{I.E. over } \hat{\pi} \\
&\lesssim k^{-s_\mu/d}\left(k^{-2s_\pi/d} + \frac{k}{n}\right)\frac{1}{n}\sum_{i=1}^n K(X_i) && \text{def of } \hat{1} \text{ \& } b_\ell.
\end{aligned}$$

Thus, Line (33) is

$$\begin{aligned}
&\lesssim k^{-2s_\mu/d}\left(k^{-2s_\pi/d} + \frac{k}{n}\right)^2 \mathbb{E}\left[\left\{\frac{1}{n}\sum_{i=1}^n K(X_i)\right\}^2\right] \\
&\lesssim k^{-2s_\mu/d}\left(k^{-2s_\pi/d} + \frac{k}{n}\right)^2 && \text{Lemma C.1.2}
\end{aligned} \tag{35}$$

As in Section C.4, Line (34) is the expected value of

$$
\begin{aligned}
&Var\left[\quad \hat{\hat{1}}\bar{\mathbb{P}}_n\left\{b_\ell K\left(\pi-\hat{\pi}\right)A\left(1-\frac{\hat{\pi}}{\pi}\right)\left(\hat{\mu}_1-\mu_1\right)\right\} \mid \mathbf{X}_{\text{all}}\right] \\
&= Var\left[\ \mathbb{E}\left[\hat{\hat{1}}\bar{\mathbb{P}}_n\left\{b_\ell K\left(\pi-\hat{\pi}\right)A\left(1-\frac{\hat{\pi}}{\pi}\right)\left(\hat{\mu}_1-\mu_1\right)\right\} \mid \mathbf{X}_{\text{all}},\hat{\pi},\hat{\mu}_1\right] \mid \mathbf{X}_{\text{all}}\right] \\
&\quad + \mathbb{E}\left[Var\left[\hat{\hat{1}}\bar{\mathbb{P}}_n\left\{b_\ell K\left(\pi-\hat{\pi}\right)A\left(1-\frac{\hat{\pi}}{\pi}\right)\left(\hat{\mu}_1-\mu_1\right)\right\} \mid \mathbf{X}_{\text{all}},\hat{\pi},\hat{\mu}_1\right] \mid \mathbf{X}_{\text{all}}\right] \qquad \text{Law of total var} \\
&= Var\left[\frac{\hat{1}}{n}\sum_{i=1}^n b_\ell K(\pi-\hat{\pi})^2(\hat{\mu}_1-\mu_1)|\mathbf{X}_{\text{all}}\right] \\
&\quad + \mathbb{E}\left[\frac{\hat{1}}{n^2}\sum_{i=1}^n b_\ell^2 K^2\left(\hat{\pi}-\pi\right)^2 Var(A|\bar{\mathbf{X}})\left(1-\frac{\hat{\pi}}{\pi}\right)^2\left(\hat{\mu}_1-\mu_1\right)^2|\mathbf{X}_{\text{all}}\right]. \\
&= \hat{1}Var\left[\frac{1}{n}\sum_{i=1}^n b_\ell(X_i)K(X_i)\epsilon_{i\pi}^2\epsilon_{i\mu}|\mathbf{X}_{\text{all}}\right] \qquad (36) \\
&\quad + \hat{1}\mathbb{E}\left[\frac{1}{n^2}\sum_{i=1}^n b_\ell(X_i)K(X_i)\epsilon_{i\pi}^2 Var(A|\bar{\mathbf{X}})\left(1-\frac{\hat{\pi}(X_i)}{\pi(X_i)}\right)^2\epsilon_{i\mu}^2|\mathbf{X}_{\text{all}}\right]. \qquad (37)
\end{aligned}
$$

Since $\hat{1}Var(\epsilon_{\hat{\pi}i}^2|\mathbf{X}_{\text{all}}) \leq ck/n$ and $\epsilon_{\hat{\pi}i}^2 \leq 1$, we can follow the same steps as in Section C.4.1 (with $(\epsilon_{\hat{\pi}i},\epsilon_{\hat{\pi}j})$ replaced throughout by $\left(\epsilon_{\hat{\pi}i}^2,\epsilon_{\hat{\pi}j}^2\right)$) to see that Line (36) has expectation $\lesssim 1/(nh^d)$. Similarly, since $\epsilon_{\hat{\pi}i}^2 \leq 1$, we can follow the same steps as in Section C.4.2 to see that Line (37) has expectation $\lesssim 1/(nh^d)$. Thus, by Markov's Inequality and Eq (35), we see that Line (31) is

$$
\begin{aligned}
&\lesssim_{\mathbb{P}} k^{-s_\mu/d}\left(k^{-2s_\pi/d}+\frac{k}{n}\right)+\sqrt{\frac{1}{nh^d}} \\
&\leq k^{-(s_\mu+s_\pi)/d}+\frac{k^{1-s_\mu/d}}{n}+\sqrt{\frac{1}{nh^d}}.
\end{aligned}
$$

C.6. Bounding Line (14) under the conditions of Point 3

If we assume only that $(\hat{\pi},\hat{\mu}_1) \perp \mathbf{Z}$, then

$$
\begin{aligned}
&\mathbb{E}\left[\hat{\hat{1}}|\bar{\mathbb{P}}_n\left\{b_\ell K\hat{\kappa}\hat{\pi}A\left(\hat{\pi}^{-1}-\pi^{-1}\right)\left(\hat{\mu}_1-\mu_1\right)\right\}| \ \middle| \ \mathbf{X}_{\text{all}}\right] \\
&\lesssim \hat{1}\bar{\mathbb{P}}_n\left\{K\,\mathbb{E}\left(\ |1-\hat{\pi}/\pi|\ |\hat{\mu}_1-\mu_1| \ \middle|\ \mathbf{X}_{\text{all}}\right)\right\} && A, b_\ell(x), \hat{\kappa}(x) \lesssim 1 \\
&\lesssim \hat{1}\bar{\mathbb{P}}_n\left\{K\,\mathbb{E}\left(\pi\,|1-\hat{\pi}/\pi|\ |\hat{\mu}_1-\mu_1| \ \mid \mathbf{X}_{\text{all}}\right)\right\} && \text{from } 1/\pi(x) \lesssim 1 \\
&\leq \hat{1}\bar{\mathbb{P}}_n\left\{K\,\mathbb{E}\left((\pi-\hat{\pi})^2|\mathbf{X}_{\text{all}}\right)^{1/2}\mathbb{E}\left(\left(\hat{\mu}_1-\mu_1\right)^2|\mathbf{X}_{\text{all}}\right)^{1/2}\right\} && \text{Cauchy Schwartz} \\
&\lesssim \left(\frac{k}{n}+k^{-2s_\mu/d}\right)^{1/2}\left(\frac{k}{n}+k^{-2s_\mu/d}\right)^{1/2}\frac{1}{n}\sum_{i=1}^n K(X_i) && (\hat{\pi},\hat{\mu}_1)\perp \mathbf{Z}, \text{ and def. of } \hat{1} \\
&\lesssim \left(\sqrt{\frac{k}{n}}+k^{-s_\mu/d}\right)\left(\sqrt{\frac{k}{n}}+k^{-s_\mu/d}\right)\frac{1}{n}\sum_{i=1}^n K(X_i) && (38) \\
&\lesssim_{\mathbb{P}} \frac{k}{n}+\frac{k^{1/2-s_\mu/d}}{\sqrt{n}}+\frac{k^{1/2-s_\pi/d}}{\sqrt{n}}+k^{-(s_\mu+s_\pi)/d} && \text{Lemma C.1.1 + Markov's Ineq.}
\end{aligned}
$$

Above, Line 38 comes from the fact that $\sqrt{a+b} \leq \sqrt{a}+\sqrt{b}$ for any two positive constants a, b.

D. Proof of Theorem 3.10

First we remark that the "reproducing" property for local polynomial estimators still holds even when $\hat{\nu}$ is pre-estimated. If f is a $\lfloor s_\tau \rfloor$ order polynomial, then there exists a set of coefficients β such that $f(x) = b(x)^\top \beta$. Thus,

$$\begin{aligned} f(x_{\text{new}}) = b(x_{\text{new}})^\top \beta =& b(x_{\text{new}})^\top \hat{\tilde{\mathbf{Q}}}^{-1} \sum_{i=1}^n b(X_i) K(X_i) \hat{\nu}(X_i) b(X_i)^\top \beta \\ =& b(x_{\text{new}})^\top \hat{\tilde{\mathbf{Q}}}^{-1} \sum_{i=1}^n b(X_i) K(X_i) \hat{\nu}(X_i) f(X_i) \\ =& \sum_{i=1}^n \hat{\tilde{w}}(X_i) f(X_i). \end{aligned} \tag{39}$$

Let $\tau_{\lfloor s_\tau \rfloor, x_{\text{new}}}$ be the $\lfloor s_\tau \rfloor$ order Taylor approximation of τ at x_{new}. It follows from Eq (39) that

$$\frac{1}{n} \sum_{i=1}^n \hat{\tilde{w}}(X_i) \tau_{\lfloor s_\tau \rfloor, x_{\text{new}}}(X_i) = \tau_{\lfloor s_\tau \rfloor, x_{\text{new}}}(x_{\text{new}}) = \tau(x_{\text{new}}), \tag{40}$$

where the second equality comes from the fact that the Taylor approximation is exact at x_{new}.

Conditional on $\hat{\nu}$ and $\bar{\mathbf{X}}$, the oracle bias is

$$\begin{aligned} &\mathbb{E}\left(\left\{ \hat{\tilde{\tau}}_{\text{oracle}}(x_{\text{new}}) - \tau(x_{\text{new}}) \right\} | \hat{\nu}, \bar{\mathbf{X}} \right) \\ &= \frac{1}{n} \sum_{i=1}^n \hat{\tilde{w}}(X_i) \mathbb{E}\left(f_{\text{DR},\theta}(Z_i) | \hat{\nu}, \bar{\mathbf{X}} \right) - \tau(x_{\text{new}}) \\ &= \frac{1}{n} \sum_{i=1}^n \hat{\tilde{w}}(X_i) \tau(X_i) - \tau(x_{\text{new}}) && \hat{\nu} \perp f_{\text{DR},\theta}(Z_i) | \bar{\mathbf{X}} \\ &= \frac{1}{n} \sum_{i=1}^n \hat{\tilde{w}}(X_i) \left\{ \tau(X_i) - \tau_{\lfloor s_\tau \rfloor, x_{\text{new}}}(X_i) \right\} && \text{Eq (40)} \\ &\leq \frac{1}{n} \sum_{i=1}^n |\hat{\tilde{w}}(X_i)| \ \left| \tau(X_i) - \tau_{\lfloor s_\tau \rfloor, x_{\text{new}}}(X_i) \right| \ |\mathcal{I}(X_i)| && \text{definitions of } \hat{\tilde{w}} \ \& \ \mathcal{I} \\ &\leq \frac{1}{n} \sum_{i=1}^n |\hat{\tilde{w}}(X_i)| \ \|X_i - x_{\text{new}}\|^{s_\tau} \ |\mathcal{I}(X_i)| && \text{Assm 3.9} \\ &\leq \frac{h^{s_\tau}}{n} \sum_{i=1}^n |\hat{\tilde{w}}(X_i)| && \text{definition of } \mathcal{I}. \end{aligned} \tag{41}$$

From here, we study the expectation of the squared oracle error multiplied by $\hat{\hat{1}}$.

$$
\begin{aligned}
&\mathbb{E}\left[\hat{\hat{1}}\left\{\hat{\hat{\tau}}_{\text{oracle}}(x_{\text{new}}) - \tau(x_{\text{new}})\right\}^2\right] \\
&= \mathbb{E}\left[\hat{\hat{1}}\mathbb{E}\left(\left\{\hat{\hat{\tau}}_{\text{oracle}}(x_{\text{new}}) - \tau(x_{\text{new}})\right\}^2 |\hat{\nu}, \bar{\mathbf{X}}\right)\right] && \text{I.E.} \\
&\leq \mathbb{E}\left[\hat{\hat{1}}\left(\frac{h^{s_\tau}}{n}\sum_{i=1}^n |\hat{\hat{w}}(X_i)|\right)^2 + \hat{\hat{1}}Var\left(\hat{\hat{\tau}}_{\text{oracle}}(x_{\text{new}})|\hat{\nu}, \bar{\mathbf{X}}\right)\right] && \text{Eq (41)} \\
&= \mathbb{E}\left[\hat{\hat{1}}\left(\frac{h^{s_\tau}}{n}\sum_{i=1}^n |\hat{\hat{w}}(X_i)|\right)^2 + \frac{\hat{\hat{1}}}{n^2}\sum_{i=1}^n \hat{\hat{w}}(X_i)^2 Var(f_{\text{DR},\theta}(Z_i)|X_i)\right] \\
&\lesssim \mathbb{E}\left[\hat{\hat{1}}\left(\frac{h^{s_\tau}}{n}\sum_{i=1}^n |\hat{\hat{w}}(X_i)|\right)^2 + \frac{\hat{\hat{1}}}{n^2}\sum_{i=1}^n \hat{\hat{w}}(X_i)^2\right] && \text{Assms 3.1 \& 3.2} \\
&= \mathbb{E}\left[\frac{\hat{\hat{1}}h^{2s_\tau}}{n^2}\sum_{i\neq j} |\hat{\hat{w}}(X_i)|\,|\hat{\hat{w}}(X_j)| + \frac{\hat{\hat{1}}(1+h^{2s_\tau})}{n^2}\sum_{i=1}^n \hat{\hat{w}}(X_i)^2\right] \\
&\lesssim \mathbb{E}\left[\frac{h^{2s_\tau}}{n^2}\sum_{i\neq j}\frac{\mathcal{I}(X_i)\mathcal{I}(X_j)}{h^{2d}} + \frac{\hat{\hat{1}}}{n^2}\sum_{i=1}^n \hat{\hat{w}}(X_i)^2\right] && \text{Lemma C.1.3} \\
&= \frac{h^{2s_\tau}}{n^2}\sum_{i\neq j}\frac{\mathbb{E}\left[\mathcal{I}(X_i)\right]\mathbb{E}\left[\mathcal{I}(X_j)\right]}{h^{2d}} + \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}\left[\hat{\hat{1}}\hat{\hat{w}}(X_i)^2\right] \\
&\lesssim \frac{h^{2s_\tau}}{n^2}\sum_{i\neq j} 1 + \frac{1}{n^2}\sum_{i=1}^n \frac{1}{h^d} && \text{Lemma C.1.1 \& C.1.5} \\
&\leq h^{2s_\tau} + \frac{1}{nh^d}.
\end{aligned}
$$

Markov's Inequality then shows the result.

E. Conditional Variance of Pseudo-outcomes

For the pseudo-outcome function $f_{\text{U},\theta}$, assume that $A \perp Y|X$ and $Var(Y|X) = \sigma^2$. It follows from $A \perp Y|X$ that $\eta(X) = \mu_1(X) = \mu_0(X)$ and $Var(Y|X, A) = Var(Y|X) = \sigma^2$. Thus, from the Law of Total Variance,

$$
\begin{aligned}
Var\left(f_{\text{U},\theta}(A, X, Y)|X\right) &= Var\left(\frac{Y - \eta(X)}{A - \pi(X)}|X\right) \\
&= \mathbb{E}\left[Var\left(\frac{Y - \eta(X)}{A - \pi(X)}|X, A\right)|X\right] + Var\left[\mathbb{E}\left(\frac{Y - \eta(X)}{A - \pi(X)}|X, A\right)|X\right] \\
&= \mathbb{E}\left[(A - \pi(X))^{-2} Var\left(Y|X, A\right)|X\right] + Var\left[\frac{\mu_A(X) - \eta(X)}{A - \pi(X)}|X\right]. \qquad (42)
\end{aligned}
$$

The assumption that $\eta(X) = \mu_A(X)$ implies that the second term in Line 42 is zero, and so

$$
\begin{aligned}
Var\left(f_{\text{U},\theta}(A, X, Y)|X\right) &= \mathbb{E}\left[(A - \pi(X))^{-2}|X\right]\sigma^2 \\
&= \left\{\frac{\pi(X)}{\{1 - \pi(X)\}^2} + \frac{1 - \pi(X)}{\{0 - \pi(X)\}^2}\right\}\sigma^2 \\
&= \left\{\frac{\pi^3 + \{1 - \pi\}^3}{(1 - \pi)^2\pi^2}\right\}\sigma^2.
\end{aligned}
$$

For $f_{\text{cov},\theta}(Z)$, if $A \perp Y|X$ and $\mathbb{E}\left[(Y-\eta(X))^2|X\right] = \sigma^2$ then

$$\begin{aligned}
Var\left(f_{\text{cov},\theta}(Z)|X\right) &= \nu(X)^{-2} Var\left[(A-\pi(X))(Y-\eta(X))|X\right] \\
&= \nu(X)^{-2}\mathbb{E}\left[(A-\pi(X))^2(Y-\eta(X))^2|X\right] \\
&\qquad - \mathbb{E}\left[(A-\pi(X))|X\right]^2 \mathbb{E}\left[(Y-\eta(X))|X\right]^2 \\
&= \nu(X)^{-2}\mathbb{E}\left[(A-\pi(X))^2|X\right]\mathbb{E}\left[(Y-\eta(X))^2|X\right] \\
&= \nu(X)^{-1}\sigma^2.
\end{aligned}$$

For $f_{\text{DR},\theta}$, if $Var(Y|A,X) = \sigma^2$ we have

$$\begin{aligned}
&Var\left(f_{\text{DR},\theta}(A,X,Y)|X\right) \\
&= Var\left[\mu_1(X) - \mu_0(X) + \frac{A-\pi(X)}{\pi(X)(1-\pi(X))}(Y-\mu_A(X))|X\right] \\
&= \nu(X)^{-2} Var\left[(A-\pi(X))(Y-\mu_A(X))|X\right] \\
&= \nu(X)^{-2}\left[Var\left\{(A-\pi(X))\mathbb{E}\left\{Y-\mu_A(X)|A,X\right\}|X\right\}\right. \\
&\qquad \left.\mathbb{E}\left\{(A-\pi(X))^2 Var\left\{Y-\mu_A(X)|A,X\right\}|X\right\}\right] \qquad \text{Law of Total Var} \\
&= \nu(X)^{-2}\left[0\right. \\
&\qquad \left.\mathbb{E}\left\{(A-\pi(X))^2|X\right\}\sigma^2\right] \\
&= \nu(X)^{-1}\sigma^2 \\
&= \kappa(X)^{-1}\pi(X)^{-1}\sigma^2.
\end{aligned}$$

F. Holder Condition Implies Local Accuracy of Taylor Expansion

For completeness, this section reviews the classic result that Holder smooth functions are close to their Taylor expansions (see, e.g., Tsybakov, 2009; Kennedy, 2023). This property follows from the fact that the residual of a Taylor expansion for a function depends on the partial derivatives of that function, which are bounded by the Holder condition.

We first introduce notation. Let $f : \mathcal{D} \to \mathbb{R}$ be a function with domain $\mathcal{D} \subset \mathbb{R}^k$. For any k-length vector of indices $\mathbf{i} = (i_1, \ldots, i_k) \in \{1, \ldots, d\}^k$ and any $x' \in \mathcal{D}$, let

$$D^{\mathbf{i}} f(x') := \left.\frac{\partial^k f(x)}{\partial x_{i_1} \ldots \partial x_{i_k}}\right|_{x=x'}$$

be the higher order partial derivative of f, over indices $\mathbf{i}$, evaluated at x'.

We say that a function f is Holder of order s if it has $\lfloor s \rfloor$ continuous derivatives, and there is a constant c such that $|D^{\mathbf{i}} f(x)| \leq c$ and

$$|D^{\mathbf{i}} f(x) - D^{\mathbf{i}} f(x')| \leq c\|x - x'\|^{s-\lfloor s \rfloor}$$

for all $x, x' \in \mathcal{D}$ and $\mathbf{i} \in \{1, \ldots, d\}^k$ with length $k \leq \lfloor s \rfloor$ (see, e.g., Belloni et al., 2015).

Lemma F.1. *If $f : \mathcal{D} \to \mathbb{R}$ is a Holder function of order s and $\mathcal{D}$ is an open subset of $\mathbb{R}^d$, then there is a constant c such that*

$$\left|f(x) - f_{\lfloor s \rfloor, x'}(x)\right| \leq c|x - x'|^s$$

for all $x, x' \in \mathcal{X}$, where $f_{\lfloor s \rfloor, x}$ is the $\lfloor s \rfloor$-order Taylor expansion of f at x'.

Proof. For $a, b, c \in \mathcal{D}$, let

$$\psi_k(a,b,c) := \frac{1}{k!}\sum_{i_1=1}^{d} \cdots \sum_{i_k=1}^{d} D^{\mathbf{i}} f(a) \prod_{j=1}^{k}\left(b_{i_j} - c_{i_j}\right)$$

be the k^{th} term in a Taylor approximation of f (see page 44 of Serfling, 1980). For $d = 1$, the function ψ_k reduces to simply

$$\psi_k(a, b, c) = \frac{1}{k!} \left. \frac{\partial^k f(t)}{\partial t^k} \right|_{t=a} (b - c)^k .$$

The n-order Taylor expansion of f at x is

$$f_{x,n}(y) = f(x) + \sum_{k=1}^{n} \psi_k(x, y, x).$$

Since f has $\lfloor s \rfloor$ continuous partial derivatives, the multivariate version of Taylor's implies that there exists a point z on the line joining x and y such that

$$\begin{aligned} f(y) &= f(x) + \left\{ \sum_{k=1}^{\lfloor s \rfloor - 1} \psi_k(x, y, x) \right\} + \psi_{\lfloor s \rfloor}(z, y, x) \\ &= f(x) + \left\{ \sum_{k=1}^{\lfloor s \rfloor} \psi_k(x, y, x) \right\} + \left\{ \psi_{\lfloor s \rfloor}(z, y, x) - \psi_{\lfloor s \rfloor}(x, y, x) \right\} \\ &= f_{\lfloor s \rfloor, x}(y) + \left\{ \psi_{\lfloor s \rfloor}(z, y, x) - \psi_{\lfloor s \rfloor}(x, y, x) \right\} \end{aligned}$$

(see page 44 of Serfling, 1980). For the univariate case, the term in braces is

$$\begin{aligned} &\frac{1}{\lfloor s \rfloor!} \left\{ \left. \frac{\partial^k f(t)}{\partial t^k} \right|_{t=z} - \left. \frac{\partial^k f(t)}{\partial t^k} \right|_{t=x} \right\} (y - x)^{\lfloor s \rfloor} \\ &\lesssim \left| \left. \frac{\partial^k f(t)}{\partial t^k} \right|_{t=z} - \left. \frac{\partial^k f(t)}{\partial t^k} \right|_{t=x} \right| |y - x|^{\lfloor s \rfloor} \\ &\lesssim |z - x|^{s - \lfloor s \rfloor} |y - x|^{\lfloor s \rfloor} && \text{by Holder condition} \\ &\leq |y - x|^{s - \lfloor s \rfloor} |y - x|^{\lfloor s \rfloor} \\ &= |y - x|^s . \end{aligned}$$

Similarly, in the multivariate case, the residual term is in braces is

$$\begin{aligned} &\frac{1}{\lfloor s \rfloor!} \sum_{i_1=1}^{d} \cdots \sum_{i_{\lfloor s \rfloor}=1}^{d} \left\{ D^{\mathbf{i}} f(z) - D^{\mathbf{i}} f(x) \right\} \prod_{j=1}^{\lfloor s \rfloor} \left(y_{i_j} - x_{i_j} \right) \\ &\lesssim \|z - x\|^{s - \lfloor s \rfloor} \sum_{i_1=1}^{d} \cdots \sum_{i_{\lfloor s \rfloor}=1}^{d} \prod_{j=1}^{i_{\lfloor s \rfloor}} \left| y_{i_j} - x_{i_j} \right| \\ &= \|z - x\|^{s - \lfloor s \rfloor} \left(\sum_{j=1}^{d} |y_j - x_j| \right)^{\lfloor s \rfloor} && (43) \\ &\leq \|z - x\|^{s - \lfloor s \rfloor} \left(\sum_{j=1}^{d} |y_j - x_j|^2 \right)^{\lfloor s \rfloor / 2} && \text{Jensen's Ineq} \\ &= \|z - x\|^{s - \lfloor s \rfloor} \|y - x\|^s \\ &\leq \|y - x\|^{s - \lfloor s \rfloor} \|y - x\|^s \\ &\leq \|y - x\|^s . \end{aligned}$$

Above, in Line (43), we use the fact that $\left(\sum_{i=1}^{n} x_i\right)^k = \sum_{i_1=1}^{n} \cdots \sum_{i_k=1}^{n} \prod_{j=1}^{k} x_{i_j}$. □

G. Sufficient Conditions for Assumption 3.6

The following lemma gives sufficient conditions for Assumption 3.6.

Lemma G.1. *(Sufficient Conditions for Assumption 3.6) Let U be a random variable uniformly distributed on the unit hypersphere, and let $\mathbf{Q}_U := \mathbb{E}\left[f_{\texttt{basis}}(U) f_{\texttt{basis}}(U)^\top\right]$. If*

1. *Assumptions 3.2, 3.4 & 3.5 hold,*
2. *$\mathbf{Q}_U$ is positive definite, and*
3. *there exists a sequence $c_n \to 0$ such that $\mathbb{E}\left\{|\hat{\nu}(x) - \nu(x)|\right\} \leq c_n$ for all x,*

then Assumption 3.6 holds.

Proof. For any symmetric matrix $\mathbf{A}$, let $\|\mathbf{A}\|$ denote the maximum absolute value of the eigenvalues of $\mathbf{A}$. Let $\bar{\mathbf{Q}}_i = b(X_i)K(X_i)\nu(X_i)b(X_i)^\top$, let $\bar{\mathbf{Q}} = \frac{1}{n}\sum_i \bar{\mathbf{Q}}_i$, and let $\mathbf{Q} = \mathbb{E}\left[\bar{\mathbf{Q}}\right]$.

We will show that $\mathbb{E}\|\hat{\bar{\mathbf{Q}}} - \bar{\mathbf{Q}}\|$ and $\mathbb{E}\|\bar{\mathbf{Q}} - \mathbf{Q}\|$ both converge to zero. These two facts will together imply that $\mathbb{E}\left[\left|\lambda_{\min}\left(\hat{\bar{\mathbf{Q}}}\right) - \lambda_{\min}\left(\mathbf{Q}\right)\right|\right]$ converges to zero, which, in turn, implies that $\left|\lambda_{\min}\left(\mathbf{Q}\right) - \lambda_{\min}\left(\hat{\bar{\mathbf{Q}}}\right)\right|$ converges in probability to zero. Finally, showing $\lambda_{\min}\left(\mathbf{Q}\right) \gtrsim \lambda_{\min}\left(\mathbf{Q}_{\mathrm{U}}\right)$ will complete the proof.

For $\mathbb{E}\|\bar{\mathbf{Q}} - \mathbf{Q}\|$, note that $\mathbb{E}[\bar{\mathbf{Q}}_i] = \mathbf{Q}$, and $\|\bar{\mathbf{Q}}_i\| \lesssim 1/h^d$ since b is bounded and $K(x) \lesssim 1/h^d$. Thus, from the Rudelson Law of Large numbers (see, e.g., Rudelson, 1999; Belloni et al., 2015, or Section 1.6.3 of Tropp, 2015),

$$\mathbb{E}\|\bar{\mathbf{Q}} - \mathbf{Q}\| \lesssim \frac{\left(1/h^d\right)\log L}{n} + \sqrt{\frac{\left(1/h^d\right)\|\mathbf{Q}\|\log L}{n}} \lesssim \frac{1}{nh^d} + \sqrt{\frac{1}{nh^d}}, \tag{44}$$

where the last $\lesssim$ comes from the fact that b, ν, and $E(K(X))$ all bounded (Lemma C.1.1), and so $\|\mathbf{Q}\|$ is bounded as well. Since $nh^d \to \infty$ by Assumption 3.5, we have $\mathbb{E}\|\bar{\mathbf{Q}} - \mathbf{Q}\| \to 0$.

For $\mathbb{E}\|\hat{\bar{\mathbf{Q}}} - \bar{\mathbf{Q}}\|$, we have

$$\begin{aligned}
\mathbb{E}\|\hat{\bar{\mathbf{Q}}} - \bar{\mathbf{Q}}\| &= \mathbb{E}\left[\max_{v:\|v\|=1}\left|\frac{1}{n}\sum_{i=1}^n \left(v^\top b(X_i)\right)^2 K(X_i)\left\{\hat{\nu}(X_i) - \nu(X_i)\right\}\right|\right] \\
&\leq \mathbb{E}\left[\max_{v:\|v\|=1}\frac{1}{n}\sum_{i=1}^n \|v\|_2^2\,\|b\|_2^2 K(X_i)\left|\hat{\nu}(X_i) - \nu(X_i)\right|\right] && \text{Cauchy Schwartz} \\
&\lesssim \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[K(X_i)\;\mathbb{E}\left\{|\hat{\nu}(X_i) - \nu(X_i)|\;|X_i\right\}\right] \\
&\leq \frac{c_n}{n}\sum_{i=1}^n \mathbb{E}\left[K(X_i)\right] \\
&\lesssim c_n,
\end{aligned} \tag{45}$$

where the last line comes from Lemma C.1.1.

Next, we show that $\lambda_{\min}\left(\mathbf{Q}\right) \gtrsim \lambda_{\min}\left(\mathbf{Q}_{\mathrm{U}}\right)$. Let $O := g(X) := \frac{X - x_{\text{new}}}{h}$, and let f_X and f_O denote the densities of X and O respectively. Note that $g^{-1}(o) = oh + x_{\text{new}}$. Let $\mathbf{J} = h\mathbf{I}$ be the Jacobian of g^{-1}, and let $|\mathbf{J}|$ be its determinant. Since g is an invertible transformation applied to a continuous variable, we have

$$f_O(o) = f_X(g^{-1}(o))\,|\mathbf{J}| \asymp |\mathbf{J}| = h^d,$$

where the $\asymp$ follows from f_X being bounded above and bounded below away from zero. Applying this, we see that for any

$v \in \mathbb{R}^d$ satisfying $\|v\| = 1$,

$$\begin{aligned} v^\top \mathbf{Q} v &= \frac{1}{h^d}\mathbb{E}\left\{\left(v^\top f_{\text{basis}}\left(\frac{X - x_{\text{new}}}{h}\right)\right)^2 \texttt{kern}\left(\frac{X - x_{\text{new}}}{h}\right)\nu(X)\right\} \\ &= \frac{1}{h^d}\mathbb{E}\left\{\left(v^\top f_{\text{basis}}\left(O\right)\right)^2 \texttt{kern}\left(O\right)\nu(g^{-1}(O))\right\} \\ &\asymp \frac{1}{h^d}\mathbb{E}\left\{\left(v^\top f_{\text{basis}}\left(O\right)\right)^2 I\left(\|O\| \le 1\right)\right\} \\ &= \frac{1}{h^d}\int_{o:\|o\|\le 1}\left(v^\top f_{\text{basis}}\left(o\right)\right)^2 f_O(o)do \\ &\asymp \int_{o:\|o\|\le 1}\left(v^\top f_{\text{basis}}\left(o\right)\right)^2 do \\ &\ge \lambda_{\min}\left(\mathbf{Q}_{\text{U}}\right). \end{aligned} \tag{46}$$

Finally, combining Eqs (44) & (45) with Lemma G.2, below, we have

$$\begin{aligned} &\mathbb{E}\left[\left|\lambda_{\min}\left(\mathbf{Q}\right) - \lambda_{\min}\left(\hat{\bar{\mathbf{Q}}}\right)\right|\right] \\ &\quad= \mathbb{E}\left[\left|\lambda_{\min}\left(\mathbf{Q}\right) - \lambda_{\min}\left(\bar{\mathbf{Q}}\right) + \lambda_{\min}\left(\bar{\mathbf{Q}}\right) - \lambda_{\min}\left(\hat{\bar{\mathbf{Q}}}\right)\right|\right] \\ &\quad\le \mathbb{E}\left[\left|\lambda_{\min}\left(\mathbf{Q}\right) - \lambda_{\min}\left(\bar{\mathbf{Q}}\right)\right| + \left|\lambda_{\min}\left(\bar{\mathbf{Q}}\right) - \lambda_{\min}\left(\hat{\bar{\mathbf{Q}}}\right)\right|\right] \\ &\quad\le \mathbb{E}\left[\|\mathbf{Q} - \bar{\mathbf{Q}}\| + \|\bar{\mathbf{Q}} - \hat{\bar{\mathbf{Q}}}\|\right] \qquad \text{Lemma G.2} \\ &\quad\lesssim c_n + \frac{1}{nh^d} + \sqrt{\frac{1}{nh^d}} \\ &\quad\to 0. \end{aligned}$$

It follows from Markov's inequality that $\mathbb{P}\left(\left|\lambda_{\min}\left(\mathbf{Q}\right) - \lambda_{\min}\left(\hat{\bar{\mathbf{Q}}}\right)\right| > \epsilon\right) \to 0$ for any ϵ. This, combined with the fact that $\lambda_{\min}\left(\mathbf{Q}\right) \gtrsim \lambda_{\min}\left(\mathbf{Q}_{\text{U}}\right) > 0$ (Eq (46)), shows the result. □

Lemma G.2. *For any symmetric, p.s.d. matrices* $\mathbf{A}$ *&* $\mathbf{B}$*, we have*

$$|\lambda_{\min}(\mathbf{A}) - \lambda_{\min}(\mathbf{B})| \le \|\mathbf{A} - \mathbf{B}\|.$$

Proof. Let $v_{\mathbf{A}}$ be the eigenvector corresponding to the smallest eigenvalue of $\mathbf{A}$ in absolute value. We consider two cases. If $\lambda_{\min}(\mathbf{B}) > \lambda_{\min}(\mathbf{A})$, then

$$\begin{aligned} |\lambda_{\min}(\mathbf{B}) - \lambda_{\min}(\mathbf{A})| &= \lambda_{\min}(\mathbf{B}) - v_{\mathbf{A}}^\top \mathbf{A} v_{\mathbf{A}} \\ &\le v_{\mathbf{A}}^\top \mathbf{B} v_{\mathbf{A}} - v_{\mathbf{A}}^\top \mathbf{A} v_{\mathbf{A}} \\ &\le \|\mathbf{B} - \mathbf{A}\|. \end{aligned}$$

If $\lambda_{\min}(\mathbf{B}) < \lambda_{\min}(\mathbf{A})$, the same reasoning shows that

$$|\lambda_{\min}(\mathbf{A}) - \lambda_{\min}(\mathbf{B})| \le \|\mathbf{A} - \mathbf{B}\|.$$

In either case, the result holds. □